AD_____

Award Number: W81XWH-10-1-0500

TITLE: Novel Prostate Cancer Pathway Modeling using Boolean Implication

PRINCIPAL INVESTIGATOR: Debashis Sahoo

CONTRACTING ORGANIZATION: Leland Stanford Junior University, Stanford, CA 94305

REPORT DATE: September 2012

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01-09-2012	Annual Summary	15 AUG 2010 - 14 AUG 2012

4. TITLE AND SUBTITLE

Novel Prostate Cancer Pathway Modeling using Boolean Implication

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-10-1-0500

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Debashis Sahoo, Jonathan R. Pollack, James D. Brooks, and Joseph Lipsick

E-Mail: sahoo@stanford.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Leland Stanford Junior University, The
Stanford, CA 94305-2004

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Purpose: Prostate cancer is the second most common cause of cancer deaths in men. **Scope**: We explore relationship between genes based on our novel approaches BooleanNet and MiDReG in prostate cancer and correlate them to patient information. Human prostate cancer is typically characterized by luminal cell expansion and the absence of basal cells. In normal prostate, tissue basal cells express Keratin 5 (KRT5) and Keratin 14 (KRT14). **Major Findings**: In the microarray datasets of primary prostate cancers, we observe a robust pattern where KRT14 high samples are always KRT5 high, but not vice versa. We summarize this in the form of a Boolean relationship: "KRT14 high => KRT5 high". We identified three groups of patients in three independent prostate cancer gene expression microarray datasets: KRT14-KRT5-, KRT14-KRT5+, and KRT14+KRT5+. Recurrence-free survival analysis of these three independent datasets revealed that KRT14-KRT5- patients have the worst, KRT14+KRT5+ patients have the best, and KRT14-KRT5+ patients have intermediate clinical outcome. Based on this data, we predict that KRT14+KRT5+ cells are upstream of KRT14-KRT5+ cells, which could be upstream of KRT14-KRT5- luminal cells in normal prostate tissue.

15. SUBJECT TERMS
Prostate, cancer, microarrays, BooleanNet, MiDReG

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	46	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Introduction

Prostate cancer is the second most common cause of cancer deaths in men. Diagnosis and pathogenesis of this disease is poorly understood. Prostate specific antigen (PSA) test is still the standard diagnostic marker for prostate cancer despite its serious limitations. Large proportions of men are being diagnosed with prostate cancer but recent studies imply that many of them don't need prostate cancer treatment. There is clearly a need for better diagnostic and prognostic marker in prostate cancer.

Recent advances in DNA microarray technology that enable the simultaneous measurement of the expression of thousands of genes in a single experiment have revolutionized current molecular biology. Already, the 21st century is witnessing an explosion in the amount of biological information on normal and disease processes. A large and exponentially growing volume of gene expression data from microarrays is now available publicly. In addition to gene expression data, massive amounts of DNA copy number data is also collected through CGH microarrays. Large amounts of high throughput genomic and epigenomic data have been collected in prostate cancer. Although these datasets have been analyzed in the literature, there are opportunities for mining these datasets in the context of all other publicly available data. High throughput genomic data shows the promise for discovery of better diagnostic and prognostic markers.

Body

Previously, we have published a novel approach to discover Boolean implications between genes using these large number of gene expression datasets. Subsequently, we used Boolean implications to successfully predict genes in B cell developmental pathway (MiDReG algorithm) [3-6]. My prostate cancer project proposes to build on our successful prediction of human B cell developmental genes which can predict pathways based on human gene expression datasets. In this report, we showed that Boolean implication predicts different state of basal cell development in normal prostate tissue. The loss of basal cell expression in cancer is correlated with the recurrence-free survival of the prostate cancer.

Boolean Implication (BooleanNet)

We downloaded 25,237 microarrays in human Affymetrix U133 Plus 2.0 platform from NCBI's GEO (Gene Expression Omnibus) database [1], and normalized using RMA (Robust Multi-chip Average) algorithm [2]. Within these datasets (with thousands of microarrays) we identified expression relationships between pairs of genes (represented by probe sets on the arrays) that follow simple "if-then" rules such as "if gene X is high, then gene Y is low," or more

simply stated: “X high ⇒ Y low” (“X high implies Y low”). In this case gene X and gene Y are rarely “high” together. We call these relationships “Boolean implications”. There are only six different types of “Boolean implications” possible in these datasets. Figure 1 outlines the six different types of Boolean implications discovered among the probe sets within the human data sets. In these scatter plots, each point represents gene X’s expression versus gene Y’s expression within an individual microarray. Each plot is divided, based on thresholds, into four quadrants: (X low, Y low), (X low, Y high), (X high, Y low), and (X high, Y high).

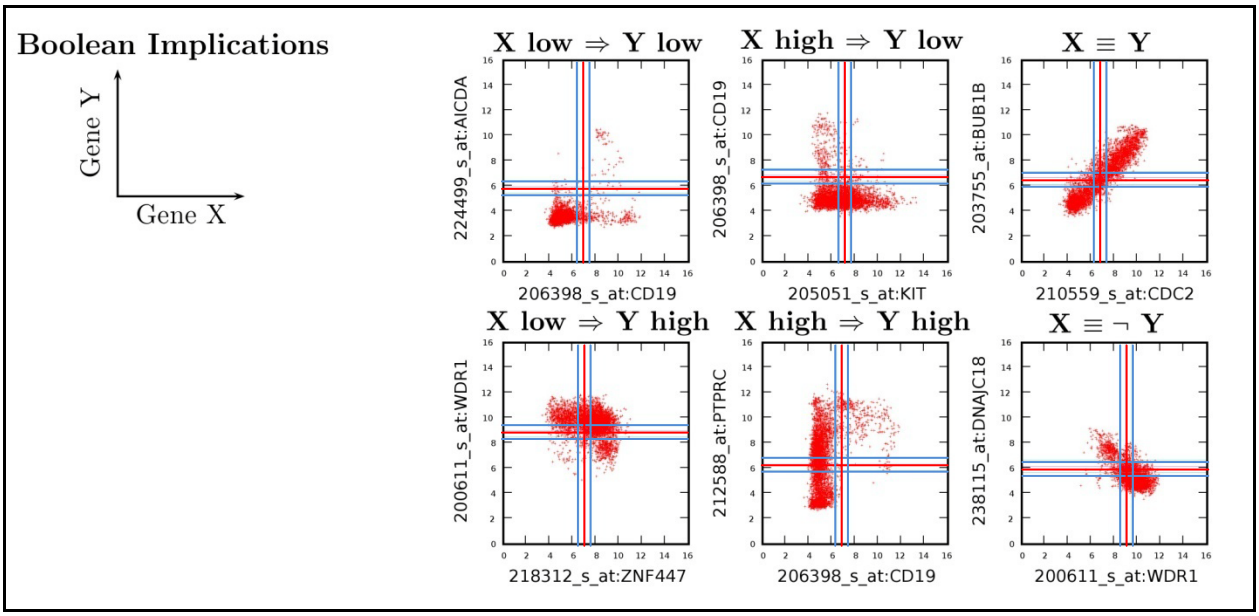


Figure 1. Boolean Implications. Scatter plots of 25,237 Affymetrix U133 Plus 2.0 human microarrays downloaded from NCBI’s Gene Expression Omnibus and normalized together. Each probeset is assigned a threshold t (red lines). Expression levels above $t + 0.5$ (blue lines) are classified as “high,” expression levels below $t - 0.5$ (blue lines) are classified as “low,” and values between $t - 0.5$ and $t + 0.5$ are classified as “intermediate.” The plots show the six different types of Boolean implication relationships between a pair of genes. Boolean implication is discovered by identifying a sparse quadrant in the scatter plot.

A Boolean implication exists when one or more quadrants is sparsely populated according to a statistical test and there are enough high and low values for each gene (to prevent the discovery of implications that follow from an extreme skew in the distribution of one of the genes) [3]. There are four asymmetric Boolean implications, each corresponding to one sparse quadrant. Two symmetric Boolean implications “equivalent” and “opposite” are discovered when two diagonally opposite sparse quadrants are identified. Boolean implications can also be extended to logical combinations of genes. For example the Boolean implication “A ⇒ B” can be discovered where A and B are either single gene conditions (e.g., X high) or logical combinations of multiple genes (e.g., X high AND Y high).

MiDReG algorithm

We developed a new method termed Mining Developmentally Regulated Genes (MiDReG) to predict genes whose expression is either activated or repressed as precursor cells differentiate [4,5]. MiDReG bases its predictions on Boolean implications mined from large-scale microarray databases and requires two or more "end point" markers for a given developmental pathway.

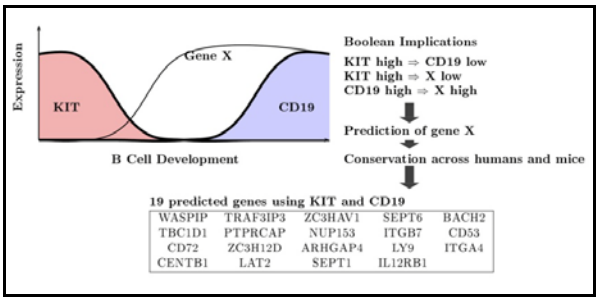


Figure 2. MiDReG algorithm.
Genes in B cell developmental pathway are discovered by using a Boolean interpolation between two known genes KIT and CD19 that marks the endpoints. KIT is expressed early in B cell development and CD19 is expressed late. There is a robust Boolean implication KIT high ⇒ CD19 low is observed in the diverse collection of microarray dataset both in humans and mice. Genes that are expressed at an intermediate step and remain high till the end are discovered by identifying genes with KIT high ⇒ X low and CD19 high ⇒ X high Boolean implications.

For example, in studies of B cell development, we used two known genes KIT and CD19 that are expressed early and late respectively during B cell development (Figure 2). A conserved Boolean implication KIT high ⇒ CD19 low is observed in the microarray dataset. MiDReG searched for genes X that are expressed during development and satisfy the implications "KIT high ⇒ X low" and "CD19 high ⇒ X high" (Figure 2), which represents the pattern of expression we expect for genes that are not expressed early in development when KIT is highly expressed (KIT high ⇒ X low), then upregulated later in development when CD19 is also upregulated (CD19 high ⇒ X high). The predicted genes were successfully validated in collaboration with the Weissman lab at Stanford University.

Novel prostate cancer pathway modeling using Boolean implication

We focused on modeling a differentiation pathway in human prostate cancer tissue using Boolean implication. This approach was motivated by our previously published MiDReG algorithm that predicts developmentally regulated genes using Boolean implication [3,4]. We first collected publicly available gene expression datasets from human prostate cancer samples (Supplementary Figure 1). To analyze the datasets using BooleanNet algorithm, we also downloaded 25,237 Affymetrix U133 Plus 2.0 datasets.

In most human epithelial tissues both Keratin 5 (K5) and Keratin 14 (K14) are expressed in the basal cell compartments. We analyzed gene expression values of K14 and K5 that is presented in the form of a scatterplot with 25,237 points representing diverse microarrays on

human samples including different normal and cancer tissues (Supplementary Figure 2). We summarize the gene expression relationship between K14 and K5 as “if K14 high then K5 high” or alternatively a Boolean implication relationship “K14 high => K5 high”. The relationship clearly suggests that K14+ arrays are a subset of K5+ arrays. Since not all cells within a sample express K14 and K5, we could hypothesize that K14+ cells are a subset of K5+ cells (Supplementary Figure 2A) based on the Boolean implication. Panel A shows a likely model of developmental gene regulation between K14 and K5, where K14 is upstream of K5 (Supplementary Figure 2).

To evaluate whether Keratin gene expression is associated with patient outcome, we investigated the status of three Keratin expression groups (KRT14+KRT5+, KRT14-KRT5+, KRT14-KRT5-) on recurrence-free survival (RFS) in three independent prostate cancer cohorts (Singh 2002 dataset, n=102; Glinsky 2004 dataset, n=78; Taylor 2010 dataset, n=185), The results confirmed that KRT14-KRT5- tumors were associated with worse clinical outcomes (B). In addition, KRT14+KRT5+ tumors were associated with best clinical and KRT14-KRT5+ tumors were associated with intermediate clinical outcome.

Training tasks

The statement of work includes several tasks on career development. I attended the 2010 Scientific Management Series from the office of postdoctoral affairs. The goals and objectives of the Scientific Management Series are to provide participants with laboratory or research management skills that will help them to launch productive independent careers in academic and other settings. All coursework was completed including Stats 141 in the Fall 2012 and the Systems Biology in spring 2011. I have been meeting with Professor Joe Lipsick weekly and Professor Jonathan Pollack biweekly. These meetings are extremely useful for my career development as I get a lot of advice on both research as well as career from both of my mentors. I have been attending the weekly seminar on Molecular Profiling Colloquium. I attend urology seminars regularly every Monday. I have been developing biological skills at the Lipsick lab and the Pollack lab. I have already acquired several biological skills such as PCR, Immunostaining to perform my own biological experiments. Overall, my training on cancer biology has been very extensive. I have already performed immunostaining on human tissues myself.

Key Research Accomplishments

1. **Collection of high-throughput genomic and epigenomic data.**

   As mentioned in my statement of work, I was planning to collect publicly available gene expression datasets for Boolean implication analysis. During last two years I have collected several publicly available gene expression datasets from National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) webpage and the Array Express webpage [1]. This collection not only includes prostate cancer but also includes gene expression data on other human cancer and normal tissues. Since all data available on a particular Affymetrix platform can be normalized together, a large database of gene expression data can be built that can be analyzed simultaneously. My largest database includes 45,000 Affymetrix microarrays. In addition to this, I have collected 23 different prostate cancer datasets in different platforms (Supplementary Figure 1). For my research work, it is important to have gene expression data annotated with clinical information such as Survival. Among the 23 different prostate cancer datasets, Survival data was available for only five different datasets (Glinsky et al., Lapointe et al., Gulzar et al., Sboner et al., Taylor et al.; Supplementary Figure 1) [10-14]. Since these datasets are in different platform, they cannot be combined together. To build a large prostate cancer specific database, I combined 14 different datasets (global prostate cancer database, total n=891) that are in Affymetrix U133A (n=456), U133A 2.0 (n=72), or U133 Plus 2.0 (n=363), selected common probesets for normalization. All these 891 samples were normalized together using standard RMA algorithm. Following are the summaries of the accomplishments.

   a. Collected 45,000 Affymetrix microarrays from NCBI's GEO

   b. Collected 23 different prostate cancer datasets

   c. Annotate five prostate cancer datasets with Survival data

   d. Combine 14 prostate cancer database to build a global prostate cancer database (n=891).

2. **Analysis of the datasets.**

   I have performed all required analysis on the collected datasets using my previously published algorithms. Following are the summaries of the accomplishments.

   a. Built a complete Boolean implication network with 45,000 Affymetrix microarrays. *I used my previously published BooleanNet algorithm (Sahoo et al. Genome Biology. 2008 [3]) on the newly collected dataset of 45,000 Affymetrix microarrays.*

   b. Identified developmental genes using MiDReG approach.

I used an approach similar to MiDReG (Mining Developmentally Regulated Genes [4,5]) to identify developmental genes in prostate tissue (Supplementary Figure 2). Human prostate cancer is typically characterized by luminal cell expansion and the absence of basal cells. In normal prostate tissue basal cell express Keratin 5 (KRT5) and Keratin 14 (KRT14). There is a significant Boolean implication between KRT5 and KRT14: "KRT14 high => KRT5 high". In other words, KRT14+ cells are a subset of KRT5+ cells. Assuming that basal cells differentiates to a luminal cell, luminal cells are predominantly KRT14- cells, and KRT14 expression change once during the development, we predict that KRT14+KRT5+ cells are upstream of KRT14-KRT5+ cells, which could be upstream of KRT14-KRT5- luminal cells in normal prostate tissue.

c. Identified correlation between developmental genes and clinical outcome.
I identified three groups of patients in three independent microarray datasets KRT14-KRT5-, KRT14-KRT5+, and KRT14+KRT5+ (Supplementary Figure 3). Recurrence free survival analysis of these three independent datasets revealed that KRT14-KRT5- patients have the worst, KRT14+KRT5+ patients have the best, and KRT14-KRT5+ patients have intermediate clinical outcome (Supplementary Figure 3). This result correlates well with the systematic loss of basal cells in prostate cancer.

3. Verify the results

My validation experiment was performed directly on human prostate tissues instead of human cell culture. We have performed KRT14 and KRT5 immunohistochemistry on 218 human prostate tissues using a tissue microarray. We discovered only 2 KRT5 positive human prostate cancer tissues and all of them were KRT14 negative. This is consistent with our hypothesis of systematic loss of basal cells in prostate cancer. Basal cells in human prostate tissue express KRT14 and KRT5 and we do not see their expression in human prostate cancer. Therefore, we believe that the correlation of KRT14 and KRT5 gene expression to recurrence-free survival must be coming from the surrounding normal human prostate tissues in prostate cancer. This is an important finding that can reveal the underlying biology of human prostate cancer.

Reportable Outcomes

1. **Abstract presentation in International Society for Stem Cell Research (ISSCR) 10th Annual Meeting, Jun 13 - 16, 2012, Yokohama, Japan. (Appendix A)**
2. **Informatics databases:**
    a. **Prostate cancer database (Supplementary Figure 1)**
    b. **Global Affymetrix gene expression database (Supplementary Figure 2)**
    c. **Bladder cancer database (Published in PNAS, Appendix B)**

d. **Colon cancer database (Submitted to NEJM)**
e. **Breast cancer database (Working draft)**
f. **Ovarian cancer database (Working draft)**
g. **Brain cancer database (Working draft)**

3. **Manuscript published:**
   a. **[Appendix B [7]] Debashis Sahoo*,** Jens-Peter Volkmer*, Robert Chin*, Philip Levy Ho, Chad Tang, Antonina V. Kurtova, Stephen B. Willingham, Senthil K. Pazhanisamy, Humberto Contreras-Trujillo, Theresa A. Storm, Yair Lotan, Andrew H. Beck, Benjamin Chung, Ash A. Alizadeh, Guilherme Godoy, Seth P. Lerner, Matt van de Rijn, Linda. D. Shortliffe, Irving L. Weissman, and Keith S. Chan. *Three differentiation states risk-stratify bladder cancer into distinct subtypes*. PNAS, 2012 Feb 7;109(6):2078-83.
   b. **[Appendix C [8]] Debashis Sahoo**[*], Piero Dalerba[*], Tomer Kalisky[*], Pradeep S. Rajendran, Mike Rothenberg, Anne A. Leyrat, Sopheak Sim, Jennifer Okamoto, John D. Johnston, Dalong Qian, Maider Zabala, Janet Bueno, Norma Neff, Jianbin Wang, Andy A. Shelton, Brendan Visser, Shigeo Hisamori, Mark van den Wetering, Hans Clevers, Michael F. Clarke[*] and Stephen R. Quake[*]. *High throughput single-cell analysis of colon tumors: biological insights and clinical applications*. Nat Biotechnol. 2011 Nov 13;29(12):1120-7.
   c. **[Appendix D [9]] Debashis Sahoo**. The power of Boolean implication networks. Front. Physio. 23 July 2012, 3:276. doi:10.3389/fphys.2012.00276 (mini review)
4. **Received NIH pathway to independence award (K99/R00) award (Appendix E).**
5. **Manuscript submitted:**
   - **Debashis Sahoo**[*], Piero Dalerba[*], Pradeep S. Rajendran, Stephen P. Miranda, Shigeo Hisamori, and Michael F. Clarke. *Gene/Protein expression predicts survival in human colon cancer*. NEJM (Under Review).
6. **Manuscript in preparation:**
   - **Debashis Sahoo**[*], Jonathan R. Pollack, Joseph Lipsick, and James D. Brooks. *Gene/Protein expression predicts survival in human prostate cancer*.

Conclusion

We showed that Boolean implication predicts different state of basal cell development in normal prostate tissue. The loss of basal cell expression in cancer is correlated with the recurrence-free survival of the prostate cancer.

References

1. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207-210 (2002).

2. Irizarry, R. A. *et al*. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).

3. Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R. & Plevritis, S. K. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol.* **9**, R157 (2008).

4. Sahoo, D., J. Seita, D. Bhattacharya, M.A. Inlay, I.L. Weissman, S.K. Plevritis, and D.L. Dill. MiDReG: A Method of Mining Developmentally Regulated Genes using Boolean Implications. Proc Natl Acad Sci U S A. 2010 Mar 30;107(13):5732-7. Epub 2010 Mar 15. PMCID: PMC2851930

5. Inlay, M.A., D. Bhattacharya, D. Sahoo, T. Serwold, J. Seita, H. Karsunky, S.K. Plevritis, D.L. Dill, and I.L.Weissman. Ly6d marks the earliest stage of B cell specification and identifies the branchpoint between B cell and T cell development. Genes And Development. 23(20):2376-81, Oct 15 2009. PMCID: PMC2764492

6. Sahoo, D., Dill, D. L., Tibshirani, R. & Plevritis, S. K. Extracting binary signals from microarray time-course data. *Nucleic Acids Res.* **35**, 3705-3712 (2007).

7. Volkmer JP, Sahoo D, Chin RK, Ho PL, Tang C, Kurtova AV, Willingham SB, Pazhanisamy SK, Contreras-Trujillo H, Storm TA, Lotan Y, Beck AH, Chung BI, Alizadeh AA, Godoy G, Lerner SP, van de Rijn M, Shortliffe LD, Weissman IL, Chan KS. *Three differentiation states risk-stratify bladder cancer into distinct subtypes*. Proc Natl Acad Sci U S A. 2012 Feb 7;109(6):2078-83. Epub 2012 Jan 19.

8. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, Sim S, Okamoto J, Johnston DM, Qian D, Zabala M, Bueno J, Neff NF, Wang J, Shelton AA, Visser B, Hisamori S, Shimono Y, van de Wetering M, Clevers H, Clarke MF, Quake SR. *Single-cell dissection of transcriptional heterogeneity in human colon tumors*. Nat Biotechnol. 2011 Nov 13; 29 (12): 1120-7

9. Sahoo, D. *The power of Boolean implication networks*. Front. Physio. 23 July 2012, 3:276. doi:10.3389/fphys.2012.00276 (mini review)

10. Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. & Sellers, W.R. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 2002: 1:203-209.

11. Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., et al. (2004). Gene

expression profiling identifies clinically relevant subtypes of prostate cancer. Proc. Natl. Acad. Sci. USA 101, 811–816.
12. Lapointe, J., Li, C., Giacomini, C.P., Salari, K., Huang, S., Wang, P., Ferrari, M.,Hernandez-Boussard, T., Brooks, J.D., and Pollack, J.R. (2007). Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. Cancer Res. 67, 8504–8510.
13. Taylor, B.S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B.S., Arora, V.K., Kaushik, P., Cerami, E., Reva, B., Antipin, Y., Mitsiades, N., Landers, T., Dolgalev, I., Major, J.E., Wilson, M., Socci, N.D., Lash, A.E., Heguy, A., Eastham, J.A., Scher, H.I., Reuter, V.E., Scardino, P.T., Sander, C., Sawyers, C.L. & Gerald, W.L.. Integrative Genomic Profiling of Human Prostate Cancer.Cancer Cell 18, 11–22, July 13, 2010
14. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest. 2004 Mar;113(6):913-23.

Appendices

Appendix A

Abstract presentation in International Society for Stem Cell Research (ISSCR) 10th Annual Meeting, Jun 13 - 16, 2012, Yokohama, Japan.

F-2247 - SYSTEMS BIOLOGY APPROACH TO STUDY STEM AND PROGENITOR CELLS OF NORMAL AND MALIGNANT HUMAN TISSUES.

Sahoo, Debashis[1], Dalerba, Piero[2], Volkmer, Jens-Peter[3], Chin, Robert K.[4], Tang, Chad[3], Willingham, Stephen B.[3], Chan, Keith S.[3], van de Rijn, Matt[1], Shortliffe, Linda D.[5], Clarke, Mike F.[3], Lipsick, Joseph[1], Weissman, Irving L.[1]

[1]Pathology, Stanford University, Stanford, CA, USA, [2]Medicine, Stanford University, Stanford, CA, USA, [3]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA, [4]Department of Radiation and Cellular Oncology, University of Chicago Medical Center, Chicago, IL, USA, [5]Urology, Stanford University, Stanford, CA, USA

Many, if not all organs and tissues consist of self-renewing stem cells that give rise to distinct, sequential progenitors with increasingly limited development potential, ultimately producing functional mature cells. All malignancies develop from cells within such hierarchies, requiring progression of events resulting in tumor cells that are capable of self-renewal, survival, migration, and likely also differentiation. The identification and characterization of stem, progenitor, and mature cells within normal and diseased tissue are not only critical for the understanding of underlying biology but also in developing more effective therapeutic strategies. Previous attempts to identify markers for cells at hierarchical stages of tissue differentiation involved either 1) large screening studies using antibody libraries or gene expression arrays, or 2) focused trials of established markers identified in other normal and diseased tissues. Unfortunately, this “random” approach is insufficient to trace complex cellular differentiation stages, and thus most often fails. Therefore a systematic approach to identify cells within tissue differentiation hierarchies is required. We applied systematic computational approaches to identify markers of stem and progenitor cells by analyzing publicly available, high-throughput gene expression datasets consisting of more than 2 billion measurement points, and subsequently to validate them using tissue microarrays. We used a new method called MiDReG (Mining Developmentally Regulated Genes) that uses Boolean implications to successfully predict genes in developmental pathways. We developed a new software tool called HEGEMON (Hierarchical Exploration of Gene Expression Microarray Online) to identify genes expressed in the stem and progenitor cells in malignant tissue development. HEGEMON explores gene expression data with its clinical information using a scatterplot of gene expression values from two genes and provides a simple framework for automatic selection of genes correlated with distinct patient information, e.g. progression and survival. Using the above tools we demonstrate a new concept that human cancers can be used as a platform to study normal developmental steps of the human tissues. We use examples of human bladder and colon cancer to show the power of this computational approach.

Appendix B

Debashis Sahoo*, Jens-Peter Volkmer*, Robert Chin*, Philip Levy Ho, Chad Tang, Antonina V. Kurtova, Stephen B. Willingham, Senthil K. Pazhanisamy, Humberto Contreras-Trujillo, Theresa A. Storm, Yair Lotan, Andrew H. Beck, Benjamin Chung, Ash A. Alizadeh, Guilherme Godoy, Seth P. Lerner, Matt van de Rijn, Linda. D. Shortliffe, Irving L. Weissman, and Keith S. Chan. *Three differentiation states risk-stratify bladder cancer into distinct subtypes*. PNAS, 2012 Feb 7;109(6):2078-83.

Three differentiation states risk-stratify bladder cancer into distinct subtypes

Jens-Peter Volkmer[a,b,c,1,2], Debashis Sahoo[a,1,2], Robert K. Chin[d,1,2], Philip Levy Ho[e], Chad Tang[a], Antonina V. Kurtova[e], Stephen B. Willingham[a], Senthil K. Pazhanisamy[e], Humberto Contreras-Trujillo[a], Theresa A. Storm[a], Yair Lotan[f], Andrew H. Beck[g], Benjamin I. Chung[b], Ash A. Alizadeh[h], Guilherme Godoy[e], Seth P. Lerner[e], Matt van de Rijn[g], Linda D. Shortliffe[b], Irving L. Weissman[a,1,2], and Keith S. Chan[e,i,1,2]

[a]Institute of Stem Cell Biology and Regenerative Medicine, [b]Department of Urology, [g]Department of Pathology, [h]Division of Hematology, and Department of Internal Medicine, Stanford University, Stanford, CA 94305; [c]Department of Urology, Heinrich Heine University, Düsseldorf, NRW 40225, Germany; [d]Department of Radiation and Cellular Oncology, University of Chicago Medical Center, Chicago, IL 60637; [f]Department of Urology, University of Texas Southwestern Medical Center, Dallas, TX 75390-9110; [e]Scott Department of Urology, and [i]Department of Molecular and Cellular Biology, Dan L. Duncan Cancer Center, Center for Cell Gene and Therapy, Baylor College of Medicine, Houston, TX 77030

Contributed by Irving L. Weissman, December 21, 2011 (sent for review November 23, 2011)

Current clinical judgment in bladder cancer (BC) relies primarily on pathological stage and grade. We investigated whether a molecular classification of tumor cell differentiation, based on a developmental biology approach, can provide additional prognostic information. Exploiting large preexisting gene-expression databases, we developed a biologically supervised computational model to predict markers that correspond with BC differentiation. To provide mechanistic insight, we assessed relative tumorigenicity and differentiation potential via xenotransplantation. We then correlated the prognostic utility of the identified markers to outcomes within gene expression and formalin-fixed paraffin-embedded (FFPE) tissue datasets. Our data indicate that BC can be subclassified into three subtypes, on the basis of their differentiation states: basal, intermediate, and differentiated, where only the most primitive tumor cell subpopulation within each subtype is capable of generating xenograft tumors and recapitulating downstream populations. We found that keratin 14 (KRT14) marks the most primitive differentiation state that precedes KRT5 and KRT20 expression. Furthermore, KRT14 expression is consistently associated with worse prognosis in both univariate and multivariate analyses. We identify here three distinct BC subtypes on the basis of their differentiation states, each harboring a unique tumor-initiating population.

Boolean analysis | stem and progenitor cells | biomarker | cancer stem cell | systems biology

Bladder cancer (BC) is the sixth most common malignancy in the United States (1), accounting for ~69,250 new cases and 14,990 deaths in 2010 (2). The vast majority (90%) of BCs are histologically classified as urothelial carcinomas (UCs) (3). UCs originate from the bladder urothelium, an epithelial tissue with a clear hierarchical organization consisting of three morphologically distinct cell types: basal, intermediate, and umbrella cells (4), representing early, mid, and later differentiation states, respectively. Malignant transformation can occur in any of these cell types thus giving rise to tumors with diverse phenotypes (5).

Currently, the World Health Organization (WHO) BC classification scheme relies primarily on pathologic stage and histological grade for prognostic classification. Identification of new molecular markers would allow for improved risk stratification so that we may better use risk-adapted therapies. Recent molecular profiling of unfractionated BCs has identified unique prognostic gene signatures (6–17). However, these gene signatures have not been clinically used and their biological relevance has remained to be elucidated. Here, we developed a biologically supervised computational approach to mine the extensive repertoire of publicly available gene expression array data to define molecular markers of cellular differentiation consistent across the range of mammalian cellular diversification (18). This algorithm uses Boolean logic to evaluate large datasets to identify genes that sequentially change expression during differentiation (e.g., progenitor genes that decrease during differentiation with the concomitant up-regulation of differentiation genes). In the current study, we have successfully predicted and functionally validated molecular markers for multiple differentiation steps in BC and analyzed their association with patient survival.

Results

In the presented study we focus on UCs, hereafter synonymously called BC, and excluded other BC subtypes (squamous and adenocarcinomas) from gene-expression, phenotypical, functional, and patient survival analyses.

Overall Strategy to Predict, Functionally Validate, and Associate Differentiation States to Survival in BC. A biologically supervised approach was used to predict markers of differentiation states in BC (Fig. 1). The expression patterns of our two previously published hierarchically related differentiation markers in BC, keratin (KRT) 5 and KRT20 (19), were analyzed by the algorithm "mining developmentally regulated genes" (MiDReG), which revealed a third differentiation marker, KRT14. We therefore hypothesized the existence of three distinct differentiation states marked by KRT14, -5, and -20, which are shared by both normal urothelium and BC. We then used the algorithm "hierarchical exploration of gene-expression microarrays online" (Hegemon) to identify cell surface markers corresponding to each differentiation state. FACS separation with these surface marker combinations allowed for the isolation of the respective tumor-initiating cell (T-IC) populations from clinical samples and analysis of their respective tumorigenic and differentiation potential in xenotransplantation models. We then analyzed the prognostic utility of these differentiation markers using patient gene-expression arrays (492 patients) and formalin-fixed paraffin-embedded (FFPE) (275 patients) tissue sets.

Author contributions: J.-P.V., D.S., R.K.C., and K.S.C. designed research; J.-P.V., D.S., R.K.C., P.L.H., C.T., A.V.K., S.B.W., S.K.P., H.C.-T., Y.L., A.H.B., G.G., S.P.L., M.v.d.R., and K.S.C. performed research; J.-P.V., D.S., R.K.C., T.A.S., B.I.C., and K.S.C. contributed new reagents/analytic tools; J.-P.V., D.S., R.K.C., A.A.A., M.v.d.R., and K.S.C. analyzed data; and J.-P.V., D.S., R.K.C., S.P.L., L.D.S., I.L.W., and K.S.C. wrote the paper.

Conflict of interest statement: I.L.W. owns Amgen stock and is a director of Stem Cells, Inc. To the authors' knowledge neither entity has a direct interest in the research reported here.

Freely available online through the PNAS open access option.

[1]J.-P.V., D.S., R.K.C., I.L.W., and K.S.C. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: jvolkmer@stanford.edu, sahoo@stanford.edu, rchin@radonc.uchicago.edu, irv@stanford.edu, or kc1@bcm.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1120605109/-/DCSupplemental.

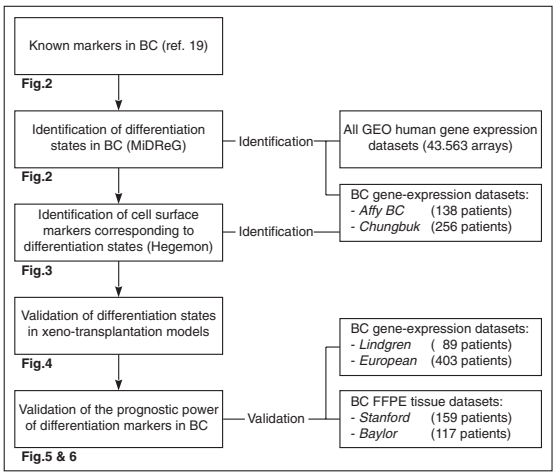


Fig. 1. Flowchart of identification and validation of differentiation states in BC. Markers of differentiation (keratins) are identified by using "mining developmentally regulated genes" (MiDReG) and corresponding cell surface markers are identified by using "hierarchical exploration of gene-expression microarray online" (Hegemon). The hypothetical hierarchy of differentiation is evaluated in patient tumor cell xenotransplantation mouse models. Association of differentiation states in bladder cancer with patient outcome is validated with existing databases and archival tissues.

Keratin 14 Is Predicted to Precede Keratin 5 and -20 in Urothelial Differentiation. KRTs are differentially expressed during epithelial tissue differentiation, a phenotype that is often conserved in neoplastic transformation (20, 21). During normal urothelial differentiation, it is proposed that basal and intermediate cells express KRT5, but not KRT20. Conversely, terminal differentiation involves the loss of KRT5 and gain of KRT20 expression (22, 23) (Fig. 2*A*). Immunofluorescence analysis confirmed our previous finding that coexpression of CD44 and KRT5 define basal/progenitor cells in BC, whereas terminally differentiated tumor cells express KRT20 but not CD44 and KRT5 (19) (Fig. S1*A*).

We developed a biologically supervised computational approach, which mines the repertoire of publicly available microarray data to identify genes that are down-regulated during cellular differentiation (18). Starting with the knowledge that KRT5 and KRT20 expression is limited to progenitor and downstream populations, respectively (Fig. 2*A*), we applied MiDReG to predict upstream keratins (KX) that satisfy two Boolean relationships (*i*) when KX expression is high, expression of progenitor KRT5 is high (Fig. 2*B*, red/blue), and (*ii*) when KX expression is high, expression of terminal differentiation marker KRT20 is low (Fig. 2*B*, red/green) (details described in *SI Methods*) (18, 24, 25). Using AffyBC and Chungbuk datasets, we identified four keratins (KRT14, KRT16, KRT6A, and KRT6B; Fig. S1*F*, details in *SI Methods*) that fulfilled these criteria (Fig. 2 *C* and *D*). Analysis of the Chungbuk dataset revealed two keratins significantly associated with outcome: KRT14 (hazard ratio (HR) 2.75, $P < 0.05$), and KRT6B (HR 3.48, $P < 0.05$) (Fig. S1*F*). We further focused on KRT14, as this marker was more highly and consistently expressed within the AffyBC and Chungbuk datasets. Immunofluoresence analysis confirmed KRT14 expression (Fig. 2*E*, green cells) marks a subpopulation of KRT5$^+$ cells in BC (Fig. 2*E*, red cells) (double positive cells, yellow, are indicated by white arrows). Analogous to BC, KRT14 staining on normal bladder tissue shows a basal-cell–restricted expression pattern (Fig. S2 *D* and *E*). On the basis of the MiDReG analysis (Fig. S2 *A*–*C*). we predicted the existence of three differentiation states in urothelial cells: basal (KRT14$^+$ KRT5$^+$KRT20$^-$), intermediate (KRT14$^-$KRT5$^+$KRT20$^-$), and differentiated (KRT14$^-$KRT5$^-$KRT20$^+$) (Fig. 2*F*).

Identification of Corresponding Surface Markers to the Predicted Keratin Differentiation States in BC. We identified surface markers specific for each of the three BC differentiation states to allow for prospective isolation by FACS and in vivo functional validation via a xenotransplantation model. To perform this analysis, we developed a software program named Hegemon (*SI Methods*) to identify surface markers highly expressed in the basal cells (KRT14$^+$) and progressively down-regulated in intermediate (KRT5$^+$) and differentiated cells (KRT20$^+$) (Fig. 3*A* and Fig. S3*F*). Using Hegemon, we ranked each marker on the basis of

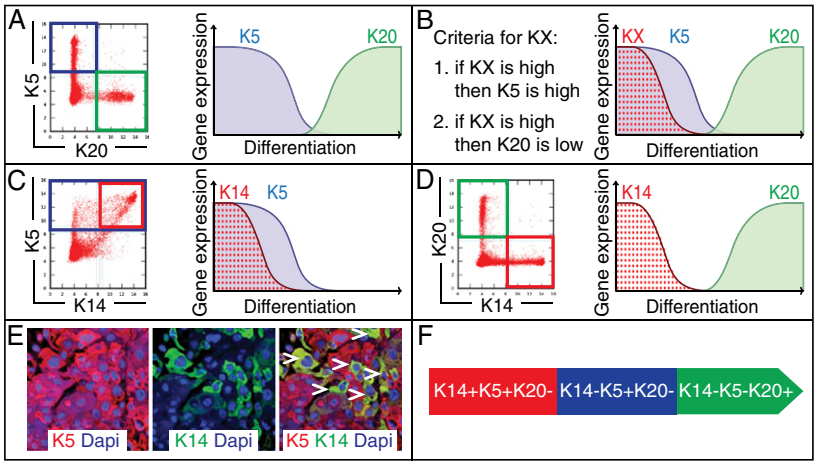


Fig. 2. Keratin 14, -5, and -20 define three differentiation states in BC. Keratins are abbreviated as KX. (*A*) K5 is expressed early during differentiation (blue) and its expression is temporally exclusive with that of the terminal differentiation marker K20 (green) in bladder cancer (BC). The mutual relationship of K5 and K20 in their temporal expression is consistent across diverse tissues (totaling 75,000 data points) in multiple species (human, mouse, and rat; Fig. S1 *B*–*E*). (*B*) Schematic illustrating the principle behind the computational strategy MiDReG used to predict a keratin X (KX, red), which is precursor to K5 and K20 by fulfilling two Boolean relationships: (*i*) when KX expression is high (red), expression of the early progenitor marker K5 is also high (blue) and (*ii*) when KX expression is high (red), expression of the differentiation marker K20 is low (green). (*C*) K14 fulfills the first Boolean relationship, its expression is high (red) when the expression of early progenitor marker K5 is also high (blue). (*D*) K14 fulfills the second Boolean relationship, its expression is low (red) when the expression of terminal differentiation marker K20 is high (green). (*E*) KRT14-expressing cells (Alexa 488/green) mark a subpopulation of KRT5$^+$ cells (Alexa 594/red) in BC; white arrows indicate double positive cells, yellow. (*F*) Schematic illustration of the three predicted differentiation states in urothelial biology.

association with patient survival (via hazard ratios) (Dataset S1) and identified CD248, S100A8, COL1A1, and CD90 (THY1) as the top candidate markers (Fig. 3*A* and Fig. S3*F* and Dataset S1). We focused on CD90, because a flow-cytometry–compatible antibody was commercially available. As expected, our previously identified marker, CD44, was also demonstrated to exhibit a predominant basal ($KRT14^+$) distribution (Fig. 3*A* and Fig. S3*F* and Dataset S1). We next used Hegemon to identify those surface markers that are expressed in all cells but down-regulated in the transition from basal to differentiated cells (Fig. 3*B* and Fig. S3*G* and Dataset S1). From this group, we focused on CD49f (ITGA6), as this marker has been reported to be coexpressed with KRT14 and is down-regulated during differentiation in various normal epithelial tissues and cancer types (26, 27).

Next, we used flow cytometry to examine whether a combination of these newly identified markers, CD90 and CD49f, and the previously identified marker CD44 could subdivide BC into distinct differentiation states. Analysis of primary tumors revealed four predicted BC populations: $CD90^+CD44^+CD49f^+$ (primitive/basal) → $CD90^-CD44^+CD49f^+$ → $CD90^-CD44^-CD49f^+$ → $CD90^-CD44^-CD49f^-$ (terminal differentiated) (Fig. 3*D*). Gene expression of KRT14, -5, and -20 in each of these purified subpopulation was analyzed by q-PCR (Fig. 3*C*). As expected, primitive/basal $CD90^+CD44^+CD49f^+$ BC cells expressed high levels of KRT14 and -5 (Fig. 3*C*, red and blue) and low levels of KRT20 (Fig. 3*C*, green). KRT14 and -5 expression were decreased in the $CD90^-CD44^+CD49f^+$ intermediate population and had the lowest expression in $CD90^-CD44^-CD49f^+$ differentiated population (Fig. 3*C*). KRT20 expression was highest in the $CD90^-CD44^-CD49f^+$ differentiated population (Fig. 3*C*).

Functional Validation of Three BC Subtypes. To functionally validate these predicted BC differentiation states, we used our unique surface marker profiles to isolate populations corresponding to each differentiation state from patient BCs using FACS (Fig. 4*A*). These isolated populations were then transplanted in vivo into immunodeficient SCID mice. As noted above, only the most upstream population harbored T-IC potential in all BCs tested. For example, in a representative BC that contained all four differentiation states (Fig. 4*A*), only the most primitive tumor cells ($CD90^+CD44^+CD49f^+$) exhibited tumorigenicity in vivo (Fig. 4*G*, basal), regenerating all downstream populations (Fig. 4*A*) and effectively reconstituting all cellular compartments from the original BC. Interestingly, within this same tumor, transplantation of a more downstream population ($CD90^-CD44^+CD49f^+$) failed to reestablish the tumor (Fig. 4*A*).

Examination of a panel of patient BC specimens revealed significant heterogeneity among tumors, some missing one or more differentiation states (Fig. 4 *C* and *E*). On the basis of our analyses, BCs could be generalized into at least three subtypes: the basal subtype, which contains all four predicted differentiation states ($CD90^+CD44^+CD49f^+$, $CD90^-CD44^+CD49f^+$, $CD90^-CD44^-CD49f^+$, and $CD90^-CD44^-CD49f^-$) (Fig. 4*A*); the intermediate subtype, which lacks the basal state (no $CD90^+CD44^+CD49f^+$ population) (Fig. 4*C*); and the differentiated subtype, which lacks both the basal and intermediate (no $CD90^+CD44^+CD49f^+$ or $CD90^-CD44^+CD49f^+$ populations) states (Fig. 4*E*). FACS isolation and subsequent xenotransplantation of sorted cells from each differentiation state from specimens representing each BC subtype revealed that only the most primitive upstream populations formed tumors (Fig. 4*G*) (e.g., in basal BC subtype, $CD90^+CD44^+CD49f^+$ cells; in intermediate BC subtype, $CD90^-CD44^+CD49f^+$ cells; and in differentiated BC subtype, $CD90^-CD44^-CD49f^+$ cells). Furthermore, the T-IC population from each BC subtype reformed only those downstream and not any upstream populations (Fig. 4 *B*, *D*, and *F*). These results revealed three phenotypically distinct BC subtypes, each containing a distinct T-IC population that invariably represented the most primitive differentiation state from that tumor (Fig. 4*G*).

Basal Subtype Is Significantly Associated with Poor Overall Survival. To evaluate the clinical significance of these three unique BC

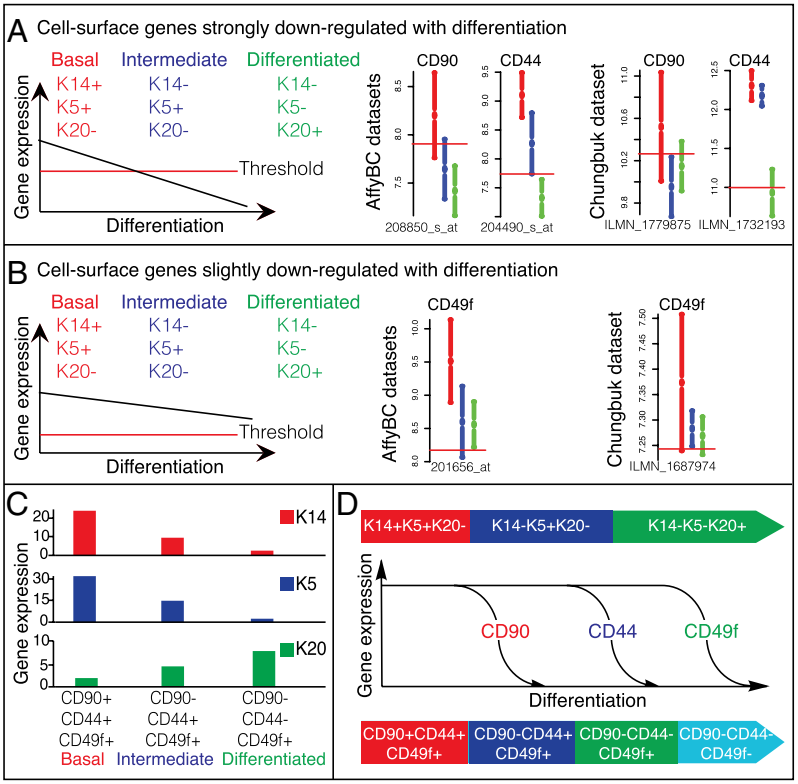
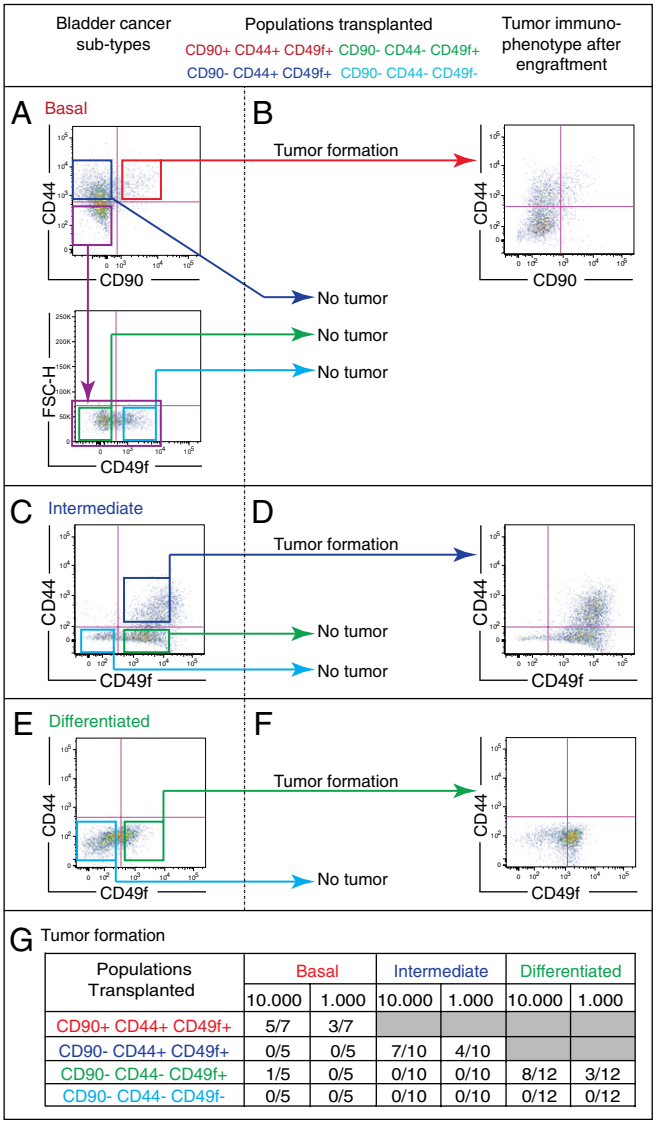


Fig. 3. Discovery of corresponding surface markers to keratins for differentiation states in BC. Keratins are abbreviated as KX. Schematics demonstrating the two criteria used to set the threshold for discovering surface markers that would correspond with the following differentiation states ($K14^+K5^+K20^-$ red; $K14^-K5^+K20^-$ blue; $K14^-K5^-K20^+$ green). The discovery analysis was performed in the AffyBC and the Chungbuk datasets (red horizontal line indicates the StepMiner-based threshold). Boxplots with mean and confidence interval for cell surface genes that fulfill the two separate criteria were shown independently. (*A*) The threshold was set in a way that would discover surface markers that were highly expressed in basal cells ($K14^+K5^+K20^-$, red) and strongly down-regulated during differentiation. (*B*) The threshold was set in a way that would discover surface markers that highly expressed all three differentiation states (red, blue, and green), and slightly down-regulated during differentiation. The detailed method of discovery and ranking of cell surface markers is presented in Fig. S3 and listed in Dataset S1. (*C*) Messenger RNA expression of K14, K5, and K20 in each of the differentiation states defined by corresponding surface markers was analyzed by real-time PCR. Corresponding surface marker combination that defines each differentiation state was listed in the *x* axis, representing basal, intermediate, and differentiated states, respectively. The relative gene-expression level was indicated in the *y* axis. (*D*) Schematic illustrating BC differentiation states as defined by keratin (K) and corresponding surface marker expression profiles.



Populations Transplanted	Basal		Intermediate		Differentiated	
	10.000	1.000	10.000	1.000	10.000	1.000
CD90+ CD44+ CD49f+	5/7	3/7				
CD90- CD44+ CD49f+	0/5	0/5	7/10	4/10		
CD90- CD44- CD49f+	1/5	0/5	0/10	0/10	8/12	3/12
CD90- CD44- CD49f-	0/5	0/5	0/10	0/10	0/12	0/12

Fig. 4. Functional validation of the computationally predicted differentiation states in BC. In vivo validation of three phenotypically distinct subtypes of bladder cancer (BC) according to their differentiation states: (*A*) basal, (*C*) intermediate, and (*E*) differentiated as defined by surface marker profiles (CD90, CD44, and CD49f). BC cells were purified by FACS and xenotransplanted intradermally into immunodeficient mice in limited dilution (10^3 and 10^4). (*B*, *D*, and *F*) The immunophenotypes of xenograft tumors derived from each BC subtype were reanalyzed by FACS postengraftment. (*A*) The basal BC subtype is composed of all differentiation states [$CD90^+CD44^+CD49f^+$ (red box) → $CD90^-CD44^+CD49f^+$ (blue box) → $CD90^-CD44^-CD49f^+$ (green box) → $CD90^-CD44^-CD49f^-$ (light blue box)]. (*B*) Only the most upstream ($CD90^+CD44^+CD49f^+$) population forms xenograft tumors and recapitulates all downstream differentiation states ($CD90^-CD44^+CD49f^+$ → $CD90^-CD44^-CD49f^+$ → $CD90^-CD44^-CD49f^-$). (*C*) The intermediate BC subtype lacks the basal differentiation state ($CD90^+CD44^+CD49f^+$). (*D*) Only the most upstream differentiation state ($CD90^-CD44^+CD49f^+$) forms xenograft tumors and can reconstitute all downstream states ($CD90^-CD44^-CD49f^+$ → $CD90^-CD44^-CD49f^-$). (*E*) In differentiated BCs that lack both the basal ($CD90^+CD44^+CD49f^+$) and the intermediate ($CD90^-CD44^+CD49f^+$) differentiation states, (*F*) only the existing differentiation state ($CD90^-CD44^-CD49f^+$) forms xenograft tumors and recapitulates the terminally differentiated ($CD90^-CD44^-CD49f^-$) downstream state. (*A*, *C*, and *E*) The terminally differentiated subpopulations ($CD90^-CD44^-CD49f^-$) never give rise to tumors. (*G*) Frequency of tumor formation of all transplanted cell populations described in *A*, *C*, and *E*.

subtypes, we analyzed their prognostic utility in two independent BC gene-expression datasets with a total of 492 patients [Lindgren $n = 89$ (9) and European $n = 403$ (6), Fig. 5]. Because the Chungbuk dataset is used as a training dataset to identify our markers, it was excluded from further validation analysis. These datasets represent all publicly available BC datasets. Survival data for two additional BC datasets were not available for analysis (10, 11). The basal BC differentiation subtype, defined by keratin ($K14^+K5^+K20^-$) or surface ($CD90^+CD44^+CD49f^+$) marker combinations were associated with worse overall survival compared with both intermediate ($K14^-K5^+K20^-$/$CD90^-CD44^+CD49f^+$) and differentiated ($K14^-K5^-K20^+$/$CD90^-CD44^-CD49f^+$) subtypes (Fig. S4 *A* and *B*). This result was additionally validated in two independent FFPE patient tissue registries with a total of 275 patients (Stanford, $n = 158$; Baylor, $n = 117$) on the basis of immunohistochemical analysis of KRT14, KRT5, and KRT20 expression (Fig. S5).

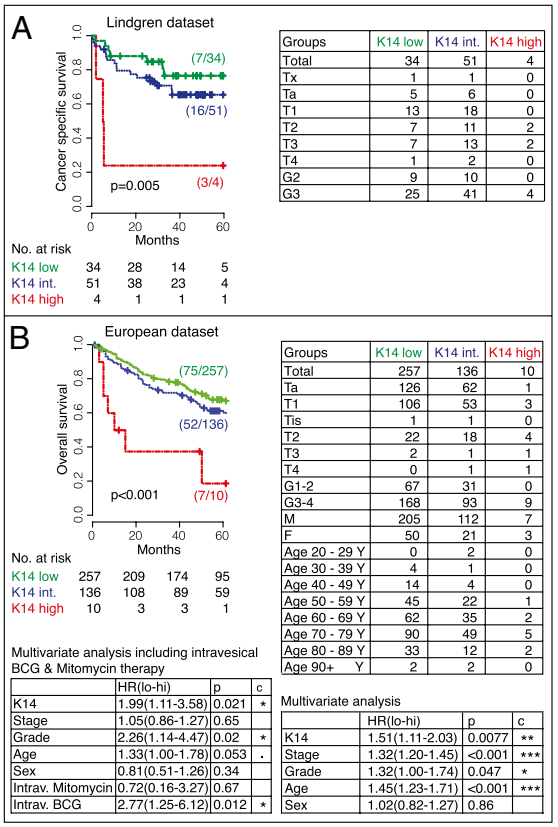
It is worth noting that a subset of patient samples in our analysis did not fit easily into one of our three BC subtypes (others, gray; Fig. S4). This additional heterogeneity may represent a block in differentiation, which may occur at any stage of differentiation. We believe that such heterogeneity complements the proposed BC differentiation states and reveals additional heterogeneity within BC subtypes that should be investigated in future analyses (Fig. S4*D*).

Basal/Primitive Cell Marker Keratin 14 Is Significantly Associated with Poor Overall Survival in BC. Because of the significantly worse overall survival associated with the basal BC subtype, and the clinical applicability of a single immunohistochemistry (IHC) marker, we evaluated the clinical significance of KRT14 as a single basal differentiation marker.

Our analysis revealed that KRT14 gene expression was associated with significantly worse overall survival in two independent datasets (Lindgren, $P = 0.005$; European, $P < 0.001$) (Fig. 5 *A* and *B*). In the European dataset, the prognostic power of KRT14 was statistically significant in both univariate and multivariate analysis when accounting for stage, grade, age, and sex (multivariate $P = 0.0077$, respectively $P = 0.021$, including tumors treated with intravesical bacillus Calmette–Guérin/chemotherapy) (Fig. 5*B*). This prognostic power remained significant when KRT14 gene expression was analyzed as a continuous variable in both uni- and multivariate analysis in the European dataset (Table S1; multivariate $P = 0.013$, respectively $P = 0.02$, including bacillus Calmette–Guérin/chemotherapy). Validation by measuring KRT14 protein expression within two independent FFPE BC tissue cohorts revealed a significant association between KRT14 and overall survival (Stanford $P < 0.0001$, multivariate $P = 0.0038$; Baylor $P = 0.009$, multivariate $P = 0.032$; Fig. 6 *A* and *B*). It is important to note that different datasets use different grading systems. Whereas the gene expression datasets are based on the 3-grade (Lindgren) or 4-grade (European) system, the FFPE BC cohorts (Stanford and Baylor) are annotated with the more recently adopted 2-grade (low and high) system. Nevertheless, the prognostic power of KRT14 holds regardless of different grading systems. Of note, although the prognostic utility of KRT14 is not confounded by pathological grade, high grade tumors are significantly enriched for KRT14 expression and vice versa (IHC datasets, Pearson's χ^2 test: Stanford, $P = 0.01$; Baylor, $P = 0.006$). Finally, subgroup analysis of clinically important BC groups, including muscle invasive disease (≥pT2), low stage disease (pTa), and patients treated with radical cystectomy, could be consistently stratified by KRT14 expression in all datasets tested (Figs. S6 and S7).

Discussion

Identification and characterization of differentiation steps are critical to our understanding of both normal tissue development and malignant transformation. During normal urothelial differentiation, it is generally accepted that basal, intermediate, and



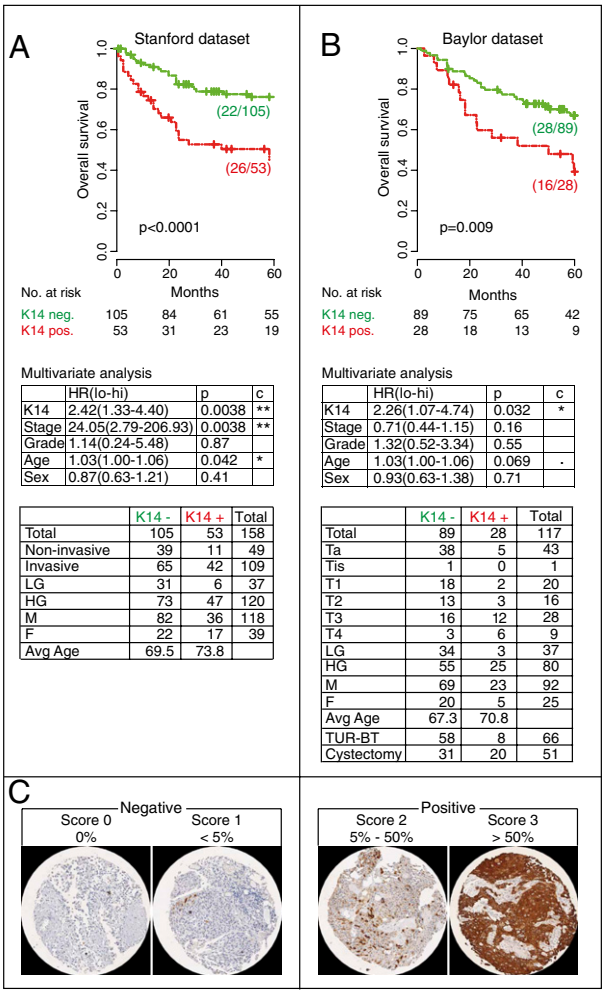
Groups	K14 low	K14 int.	K14 high
Total	34	51	4
Tx	1	1	0
Ta	5	6	0
T1	13	18	0
T2	7	11	2
T3	7	13	2
T4	1	2	0
G2	9	10	0
G3	25	41	4

Groups	K14 low	K14 int.	K14 high
Total	257	136	10
Ta	126	62	1
T1	106	53	3
Tis	1	1	0
T2	22	18	4
T3	2	1	1
T4	0	1	1
G1-2	67	31	0
G3-4	168	93	9
M	205	112	7
F	50	21	3
Age 20 - 29 Y	0	2	0
Age 30 - 39 Y	4	1	0
Age 40 - 49 Y	14	4	0
Age 50 - 59 Y	45	22	1
Age 60 - 69 Y	62	35	2
Age 70 - 79 Y	90	49	5
Age 80 - 89 Y	33	12	2
Age 90+ Y	2	2	0

	HR(lo-hi)	p	c
K14	1.99(1.11-3.58)	0.021	*
Stage	1.05(0.86-1.27)	0.65	
Grade	2.26(1.14-4.47)	0.02	*
Age	1.33(1.00-1.78)	0.053	.
Sex	0.81(0.51-1.26)	0.34	
Intrav. Mitomycin	0.72(0.16-3.27)	0.67	
Intrav. BCG	2.77(1.25-6.12)	0.012	*

	HR(lo-hi)	p	c
K14	1.51(1.11-2.03)	0.0077	**
Stage	1.32(1.20-1.45)	<0.001	***
Grade	1.32(1.00-1.74)	0.047	*
Age	1.45(1.23-1.71)	<0.001	***
Sex	1.02(0.82-1.27)	0.86	

Fig. 5. Keratin 14 gene expression is associated with worse patient survival in BC. Kaplan–Meier analysis of the probability of cancer-specific (*A*) and overall (*B*) survival according to differentiation states in bladder cancer as defined by Keratin 14 (K14) gene-expression level in two independent datasets, Lindgren (*A*) and European (*B*).



	HR(lo-hi)	p	c
K14	2.42(1.33-4.40)	0.0038	**
Stage	24.05(2.79-206.93)	0.0038	**
Grade	1.14(0.24-5.48)	0.87	
Age	1.03(1.00-1.06)	0.042	*
Sex	0.87(0.63-1.21)	0.41	

	HR(lo-hi)	p	c
K14	2.26(1.07-4.74)	0.032	*
Stage	0.71(0.44-1.15)	0.16	
Grade	1.32(0.52-3.34)	0.55	
Age	1.03(1.00-1.06)	0.069	.
Sex	0.93(0.63-1.38)	0.71	

	K14 -	K14 +	Total
Total	105	53	158
Non-invasive	39	11	49
Invasive	65	42	109
LG	31	6	37
HG	73	47	120
M	82	36	118
F	22	17	39
Avg Age	69.5	73.8	

	K14 -	K14 +	Total
Total	89	28	117
Ta	38	5	43
Tis	1	0	1
T1	18	2	20
T2	13	3	16
T3	16	12	28
T4	3	6	9
LG	34	3	37
HG	55	25	80
M	69	23	92
F	20	5	25
Avg Age	67.3	70.8	
TUR-BT	58	8	66
Cystectomy	31	20	51

Fig. 6. Keratin 14 protein expression is associated with worse patient survival in BC. (*A* and *B*) Kaplan–Meier analysis of the probability of overall survival according to differentiation states in bladder cancer as defined by keratin 14 (K14) in two independent tissue datasets, Stanford (*A*) and Baylor (*B*). (*C*) Representative micrographs of K14 IHC staining, scoring (0–3), and stratification (negative, 0–1; positive, 2–3) are presented.

umbrella cells represent sequential differentiation, from primitive to mature. It is likely that malignant transformation can occur in any of these cell types to form tumors with distinct T-IC populations (5). Our results indicate a multistep differentiation hierarchy in BCs that parallels normal urothelial differentiation. The resulting unique classification scheme broadly divides BC into three differentiation subtypes—basal, intermediate, and differentiated. We further demonstrated that each BC subtype possesses its own phenotypically distinct T-IC population within its most primitive compartment. Such a T-IC population exists at the top of a hierarchical relationship and is capable of reconstituting all downstream populations. These results add complexity to our originally proposed T-IC model (19) and suggest BC conforms to the cancer stem cell model (19, 28–33).

A subset of patient samples in our analysis does not fit into the three BC subtypes, which may reflect additional diversity. However, we did not find evidence of cellular plasticity as recently described by Chaffer et al. (34). In our functional in vivo studies, BC cells give rise to downstream differentiation states but are incapable of reforming upstream populations. More stringent biological assays such as lineage tracing in mice can be explored in future to provide definitive evidence supporting our hierarchy model.

Stratification of patients by BC subtypes, using keratin and cell surface markers, showed significant prognostic utility. Moreover, KRT14 expression is strongly associated with poor survival, independent of established clinical and pathological variables including stage, grade, age, and sex. For example, KRT14 identifies patients with worse outcome in both nonmuscle invasive (pTa) and muscle invasive (pT2 and ≥pT2) tumors. Within the muscle invasive cohort, identification of high-risk patients may allow for effective early utilization of aggressive therapies like neoadjuvant chemotherapy and provide another means to stratify patients in clinical trials. These considerations provide strong rationale for prospective studies evaluating KRT14 expression as a risk-stratifying marker.

The prognostic utility of KRT14 held when tumors were analyzed by both gene expression and IHC, the latter being a technique easily added to the repertoire of clinical laboratories. However, our IHC analysis identified relatively more KRT14 positive patients than gene-expression analysis. There are two possible explanations: differences in patient cohorts and assay sensitivity. The IHC data were obtained from patients treated at Stanford University and Baylor College of Medicine, which are tertiary referral centers that commonly treat advanced-stage BCs (59% with invasive disease), whereas gene-expression data were obtained from patients treated at multiple different European centers (ranging from primary to tertiary centers) and therefore had overall less advanced BCs (19% with invasive disease). Additionally, gene-expression analysis averages mRNA expression throughout an entire sample, whereas IHC provides resolution up to a single cell. Therefore, the same patient who may appear

KRT14 negative in a gene-expression analysis may be identified through IHC as KRT14 positive. However, the fact that both gene-expression and IHC analyses indicated KRT14 as an independent prognostic marker speaks to the robustness of this early progenitor cell marker in BC prognosis.

In addition to the differences between gene expression and IHC analysis, the nature of a retrospective study has its own limitations. For example, important clinical parameters such as lymph node status, detailed cytopathological features, and full treatment history are not always available. Additionally, the distribution of clinicopathological features in the study cohorts may not reflect the natural patient distribution. For example, carcinoma in situ cases are relatively underrepresented in all of the datasets used in this study. To overcome these limitations, the clinical utility of KRT14 needs to be validated in future prospective trials.

In summary, we have developed a unique computational strategy to identify prognostic markers linked to cellular differentiation. We subsequently validated a set of distinct differentiation markers in BC through in vivo assays and clinical outcomes analyses. It is likely that this method can be readily generalizable to other cancers. Our results hold immediate implications to understanding BC biology and further development of unique targeted therapies. Finally, our analysis revealed a clinically applicable marker, KRT14, which we believe is an ideal candidate for a large prospective trial to assess risk-adapted therapies.

Methods

Data Collection, Processing, and Statistical Analysis. See *SI Methods* for further details.

Identification of Differentiation States Using MiDReG and Identification of Corresponding Surface Markers using Hegemon. See Figs. S1 and S3, Table S1, and *SI Methods* for further details.

Immunofluorescence Staining and Immunohistochemistry. See *SI Methods* for further details.

Bladder Tumor Tissue Dissociation, Flow Cytometry Analysis and Cell Sorting, and Xenografting. Dissociation, FACS analysis, sorting, and xenografting were preformed as previously described (19). See *SI Methods* for further details.

Patient Classification for Outcome Analysis According to Bladder Cancer Differentiation Status. See *SI Methods* for further details.

ACKNOWLEDGMENTS. We thank J. Lipsick, F. Scheeren, P. Dalerba, S. Mitra, M. Diehn, S. Hilsenbeck, K. C. Osborne, D. Rowley, J. Rosen, J. D. Brooks, and R. Levy for critical discussion, helpful suggestions, and technical advice; J. Liao, H. Gill, J. Presti, M. Gonzalgo, J. Bruno, and J. Santos for consenting patients and providing bladder cancer specimens for the current study; L. Jerabek and A. Mosley for laboratory and mouse management; and K. Montgomery, S. Varma, W. Jian, and R.Ashfaq for tissue sectioning, staining, and scanning. J-P.V. is supported by Deutsche Forschungsgemeinschaft Grant VO 1704/1–1 and grants from the Urologisch Wissenschaftliche Gesellschaft. D.S. is supported by National Institutes of Health (NIH) Grant K99CA151673-01A1, Department of Defense Grant W81XWH-10-1-0500, and a grant from the Siebel Stem Cell Institute and the Thomas and Stacey Siebel Foundation. R.K.C. is supported by Radiological Society of North America Grants RR0832 and RR0907. C.T. is supported by the Howard Hughes Medical Institute Medical Fellow Program and the Stanford Medical Scholar Program. S.B.W. is supported by a grant from the Lacob Program of Excellence in Gynecologic-Ovarian Cancer Research and Treatment. H.C.-T. is supported by California Institute of Regerenative Medicine Grant TB1-01190. I.L.W. is supported by a Ludwig Institute Grant, the Jim and Carolyn Pride Family Fund, the Smith Family Foundation, and NIH Grant P01CA139490. K.S.C. is supported by the National Cancer Institute Grant R00CA129640-04 and the V Foundation for Cancer Research V Scholar Award.

1. US Cancer Statistics Working Group (2010) United States Cancer Statistics: 1999–2007 Incidence and Mortality Web-Based Report (US Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute, Atlanta). Available at http://apps.nccd.cdc.gov/uscs/toptencancers.aspx. Accessed November 9, 2010.
2. Jemal A, Siegel R, Xu J, Ward E (2010) Cancer statistics, 2010. *CA Cancer J Clin* 60: 277–300.
3. Wu XR (2005) Urothelial tumorigenesis: A tale of divergent pathways. *Nat Rev Cancer* 5:713–725.
4. Lewis SA (2000) Everything you wanted to know about the bladder epithelium but were afraid to ask. *Am J Physiol Renal Physiol* 278:F867–F874.
5. Weissman I (2005) Stem cell research: Paths to cancer therapies and regenerative medicine. *JAMA* 294:1359–1366.
6. Dyrskjøt L, et al. (2007) Gene expression signatures predict outcome in non-muscle-invasive bladder carcinoma: A multicenter validation study. *Clin Cancer Res* 13: 3545–3551.
7. Kim WJ, et al. (2010) Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer. *Mol Cancer* 9:3.
8. Kim WJ, et al. (2011) A four-gene signature predicts disease progression in muscle invasive bladder cancer. *Mol Med* 17:478–485.
9. Lindgren D, et al. (2010) Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome. *Cancer Res* 70:3463–3472.
10. Sanchez-Carbayo M, Socci ND, Lozano J, Saint F, Cordon-Cardo C (2006) Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays. *J Clin Oncol* 24:778–789.
11. Mitra AP, et al. (2009) Generation of a concise gene panel for outcome prediction in urinary bladder cancer. *J Clin Oncol* 27:3929–3937.
12. Smith SC, et al. (2011) A 20-gene model for molecular nodal staging of bladder cancer: Development and prospective assessment. *Lancet Oncol* 12:137–143.
13. Monzon FA, et al. (2009) Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. *J Clin Oncol* 27:2503–2508.
14. Stransky N, et al. (2006) Regional copy number-independent deregulation of transcription in cancer. *Nat Genet* 38:1386–1396.
15. Wild PJ, et al. (2005) Gene expression profiling of progressive papillary noninvasive carcinomas of the urinary bladder. *Clin Cancer Res* 11:4415–4429.
16. Modlich O, Prisack HB, Munnes M, Audretsch W, Bojar H (2004) Immediate gene expression changes after the first course of neoadjuvant chemotherapy in patients with primary breast cancer disease. *Clin Cancer Res* 10:6418–6431.
17. Karni-Schmidt O, et al. (2011) Distinct expression profiles of p63 variants during urothelial development and bladder cancer progression. *Am J Pathol* 178:1350–1360.
18. Sahoo D, et al. (2010) MiDReG: A method of mining developmentally regulated genes using Boolean implications. *Proc Natl Acad Sci USA* 107:5732–5737.
19. Chan KS, et al. (2009) Identification, molecular characterization, clinical prognosis, and therapeutic targeting of human bladder tumor-initiating cells. *Proc Natl Acad Sci USA* 106:14016–14021.
20. Fuchs E (1993) Epidermal differentiation and keratin gene expression. *J Cell Sci Suppl* 17:197–208.
21. Chu PG, Weiss LM (2002) Keratin expression in human tissues and neoplasms. *Histopathology* 40:403–439.
22. Johansson SL, Cohen SM (1997) Epidemiology and etiology of bladder cancer. *Semin Surg Oncol* 13:291–298.
23. De La Rosette J, Smedts F, Schoots C, Hoek H, Laguna P (2002) Changing patterns of keratin expression could be associated with functional maturation of the developing human bladder. *J Urol* 168:709–717.
24. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK (2008) Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol* 9:R157.
25. Inlay MA, et al. (2009) Ly6d marks the earliest stage of B-cell specification and identifies the branchpoint between B-cell and T-cell development. *Genes Dev* 23:2376–2381.
26. Stingl J, et al. (2006) Purification and unique properties of mammary epithelial stem cells. *Nature* 439:993–997.
27. Lim E, et al. (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat Med* 15:907–913.
28. He X, et al. (2009) Differentiation of a highly tumorigenic basal cell compartment in urothelial carcinoma. *Stem Cells* 27:1487–1495.
29. Yang YM, Chang JW (2008) Bladder cancer initiating cells (BCICs) are among EMA-CD44v6+ subset: Novel methods for isolating undetermined cancer stem (initiating) cells. *Cancer Invest* 26:725–733.
30. She JJ, Zhang PG, Wang ZM, Gan WM, Che XM (2008) Identification of side population cells from bladder cancer cells by DyeCycle Violet staining. *Cancer Biol Ther* 7: 1663–1668.
31. Su Y, et al. (2010) Aldehyde dehydrogenase 1 A1-positive cell population is enriched in tumor-initiating cells and associated with progression of bladder cancer. *Cancer Epidemiol Biomarkers Prev* 19:327–337.
32. Chan KS, Volkmer JP, Weissman I (2010) Cancer stem cells in bladder cancer: A revisited and evolving concept. *Curr Opin Urol* 20:393–397.
33. Reya T, Morrison SJ, Clarke MF, Weissman IL (2001) Stem cells, cancer, and cancer stem cells. *Nature* 414:105–111.
34. Chaffer CL, et al. (2011) Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. *Proc Natl Acad Sci USA* 108:7950–7955.

Correction

MEDICAL SCIENCES, COMPUTER SCIENCES

Correction for "Three differentiation states risk-stratify bladder cancer into distinct subtypes," by Jens-Peter Volkmer, Debashis Sahoo, Robert K. Chin, Philip Levy Ho, Chad Tang, Antonina V. Kurtova, Stephen B. Willingham, Senthil K. Pazhanisamy, Humberto Contreras-Trujillo, Theresa A. Storm, Yair Lotan, Andrew H. Beck, Benjamin I. Chung, Ash A. Alizadeh, Guilherme Godoy, Seth P. Lerner, Matt van de Rijn, Linda D. Shortliffe, Irving L. Weissman, and Keith S. Chan, which appeared in issue 6, February 7, 2012, of *Proc Natl Acad Sci USA* (109:2078–2083; first published January 19, 2012; 10.1073/pnas.1120605109).

The authors note that Robert K. Chin should be listed as an additional corresponding author. The corrected correspondence footnote appears below. The online version has been corrected.

[1]To whom correspondence may be addressed. E-mail: jvolkmer@stanford.edu, sahoo@stanford.edu, rchin@radonc.uchicago.edu, irv@stanford.edu, or kc1@bcm.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1201493109

Appendix C

Debashis Sahoo[*], Piero Dalerba[*], Tomer Kalisky[*], Pradeep S. Rajendran, Mike Rothenberg, Anne A. Leyrat, Sopheak Sim, Jennifer Okamoto, John D. Johnston, Dalong Qian, Maider Zabala, Janet Bueno, Norma Neff, Jianbin Wang, Andy A. Shelton, Brendan Visser, Shigeo Hisamori, Mark van den Wetering, Hans Clevers, Michael F. Clarke[*] and Stephen R. Quake[*]. *High throughput single-cell analysis of colon tumors: biological insights and clinical applications*. Nat Biotechnol. 2011 Nov 13;29(12):1120-7.

nature biotechnology

Single-cell dissection of transcriptional heterogeneity in human colon tumors

Piero Dalerba[1,2,9], Tomer Kalisky[3,9], Debashis Sahoo[1,9], Pradeep S Rajendran[1], Michael E Rothenberg[1,4], Anne A Leyrat[3], Sopheak Sim[1], Jennifer Okamoto[3,5], Darius M Johnston[1,3,5], Dalong Qian[1], Maider Zabala[1], Janet Bueno[6], Norma F Neff[3], Jianbin Wang[3], Andrew A Shelton[7], Brendan Visser[7], Shigeo Hisamori[1], Yohei Shimono[1], Marc van de Wetering[8], Hans Clevers[8], Michael F Clarke[1,2,9] & Stephen R Quake[3,5,9]

Cancer is often viewed as a caricature of normal developmental processes, but the extent to which its cellular heterogeneity truly recapitulates multilineage differentiation processes of normal tissues remains unknown. Here we implement single-cell PCR gene-expression analysis to dissect the cellular composition of primary human normal colon and colon cancer epithelia. We show that human colon cancer tissues contain distinct cell populations whose transcriptional identities mirror those of the different cellular lineages of normal colon. By creating monoclonal tumor xenografts from injection of a single ($n = 1$) cell, we demonstrate that the transcriptional diversity of cancer tissues is largely explained by *in vivo* multilineage differentiation and not only by clonal genetic heterogeneity. Finally, we show that the different gene-expression programs linked to multilineage differentiation are strongly associated with patient survival. We develop two-gene classifier systems (*KRT20* versus *CA1*, *MS4A12*, *CD177*, *SLC26A3*) that predict clinical outcomes with hazard ratios superior to those of pathological grade and comparable to those of microarray-derived multigene expression signatures.

The *in vivo* cellular composition of solid tissues is often difficult to investigate in a comprehensive and quantitative way. Techniques such as immunohistochemistry and flow cytometry are limited by the availability of antigen-specific monoclonal antibodies and by the small number of parallel measurements that can be performed on each individual cell. Traditional high-throughput assays, such as gene-expression arrays, when performed on whole tissues, provide information on average gene expression levels, and can be correlated only indirectly to quantitative modifications in cellular subpopulations. These limitations become particularly difficult to overcome when studying minority populations, such as stem cells, whose identification is made elusive by their low numbers and by the lack of exclusive markers. Moreover, in pathological states, such as cancer, it is usually impossible to determine whether perturbations in gene expression detected in whole tissues are due to modifications in the relative composition of different cell types or to aberrations in the gene-expression profile of mutated cells.

For example, although it has been postulated that multilineage differentiation can contribute to tumor heterogeneity[1–3], this issue remains controversial[4]. Many in the field view cancer heterogeneity mainly as the result of clonal evolution secondary to genomic instability[5,6]. Previous studies addressed this question, but could rely only on *in vitro* cultured cell lines and on simple morphological evidence[7–9]. Moreover, recent evidence indicates that, in the absence of a molecular proof of monoclonal origin, results from *in vitro* experiments based on limiting dilution can be biased due to a dramatic increase in cell survival by cell hetero-doublets. This phenomenon is best exemplified in the case of the mouse small intestine, where growth and expansion of $LGR5^+$ progenitor cells is dramatically enhanced by the presence of bystander epithelial feeder cells[10]. Based on these studies, it remained difficult to perform a quantitative measure of the degree of multilineage differentiation in cancer tissues and, above all, to investigate to what degree it actually translated into the differential activation of distinct transcriptional programs that would mirror and recapitulate the physiological processes observed in normal tissues. In this study we developed a method to dissect and investigate at the single-cell level the gene-expression profile of the distinct cell populations contained in primary human colon epithelia, both normal and neoplastic.

RESULTS

Description and technical validation of single-cell PCR

We combined fluorescence activated cell sorting (FACS) and single-cell PCR gene-expression analysis to perform a high-throughput transcriptional analysis of the distinct cellular populations contained in solid human tissues (**Supplementary Figs. 1** and **2**). This method exploits the capacity of modern flow cytometers to sort

[1]Stanford Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California, USA. [2]Department of Medicine, Division of Oncology, Stanford University, Stanford, California, USA. [3]Department of Bioengineering, Stanford University, Stanford, California, USA. [4]Department of Medicine, Division of Gastroenterology and Hepatology, Stanford University, Stanford, California, USA. [5]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [6]Tissue Bank, Stanford University, Stanford, California, USA. [7]Department of Surgery, Stanford University, Stanford, California, USA. [8]Hubrecht Institute for Developmental Biology and Stem Cell Research, Utrecht, The Netherlands. [9]These authors contributed equally to this work. Correspondence should be addressed to S.R.Q. (quake@stanford.edu) or M.F.C.(mfclarke@stanford.edu).

Received 2 May; accepted 12 October; published online 13 November 2011; doi:10.1038/nbt.2038

individual single cells with accuracy and precision (**Supplementary Fig. 3**), together with the use of microfluidic technologies to perform high-sensitivity multiplexed PCR from minute amounts of mRNA, thereby allowing parallel analysis of the expression of up to 96 genes for each individual cell. The large number of measurements per cell and the possibility to analyze several hundred cells in parallel from the same sample allow the use of statistical clustering algorithms to associate cells with similar gene expression profiles into well-defined subpopulations (**Supplementary Fig. 2**). Microfluidic platforms have been previously validated for single-cell gene-expression analysis[11–13]. Consistent with those results, our control experiments with titrated mRNA standards as well as single-cell experiments on a cell line validated the sensitivity of this approach for high-throughput analysis across multiple genes (**Supplementary Fig. 4**).

Analysis of normal human colon epithelium

We first applied single-cell PCR to the study of normal human colon epithelial cells. Human colon epithelium is composed of heterogeneous populations of cells that express different protein markers based on their lineage, differentiation stage and functional status. Many of these cell subsets can be identified by immunohistochemistry against well-characterized markers, such as MUC2, expressed by goblet cells; MKI67, expressed by proliferating cells; KRT20 and CEACAM1 (also known as CD66a), preferentially expressed by cells at the top of the colonic crypt (**Fig. 1a–d**)[14].

Under normal conditions, immature colon epithelial cells reside at the bottom of the colonic crypts (bottom-of-the-crypt cells) and express high levels of the surface marker CD44, whereas differentiated mature cells progressively migrate to the top (top-of-the-crypt cells) and progressively lose CD44 expression[14,15]. We focused our analysis on the stem and progenitor cell compartments of the colonic epithelium by sorting the EpCAM^{high}/CD44^{+} population (**Fig. 1e**,**f**; P12) which, in normal tissues, corresponds to the bottom of the human colonic crypt[14]. To study the more mature, terminally differentiated cell populations, we sorted and analyzed an equal number of cells from the EpCAM^{+}/CD44^{-}/CD66a^{high} population, which corresponds to the top of the human colonic crypt (**Fig. 1e**,**f**; P11)[16].

We first tested the ability of single-cell PCR gene-expression analysis to distinguish different cell populations using well-established reference markers. We analyzed and clustered colon epithelial cells using three genes encoding markers linked to either one of the two major cell lineages (that is, *MUC2* for goblet cells and *CA1* for enterocytes) or the immature compartment (that is, *LGR5*) of the colon epithelium[14,17–19]. This experiment showed that genes encoding lineage-specific markers are frequently expressed in a mutually exclusive way, mirroring the expression pattern of corresponding proteins (**Supplementary Fig. 5**).

We then searched for gene-expression markers of the different cell populations, with a special focus on putative stem cell markers. We mined 1,568 publicly available gene-expression array data sets from human colon epithelia (**Supplementary Table 1**), using a bioinformatics approach designed to identify developmentally regulated genes based on Boolean implication logic (**Supplementary Fig. 6**)[20]. The search yielded candidate genes whose expression was associated with that of other markers previously linked to individual colon epithelial cell lineages (**Supplementary Figs. 7–9**). Using an iterative approach, we screened >230 genes on eight independent samples of normal human colon epithelium by single-cell PCR gene-expression analysis. At each round, genes that were noninformative (that is, not differentially expressed in either positive or negative association with *CA1*, *MUC2* or *LGR5*) were removed and replaced with new candidate genes. Thereby, we progressively built a list of 57 TaqMan assays that allowed us to analyze the expression pattern of 53 distinct genes (3 housekeeping, 3 proliferation-related and 47 differentially expressed genes; **Supplementary Table 2**) with high robustness (**Supplementary Fig. 10**). This allowed us to characterize multiple cell populations, using both hierarchical clustering (**Fig. 1g**) and principal component analysis (PCA; **Fig. 1h**,**i**).

Analysis of the EpCAM^{+}/CD44^{-}/CD66a^{high} population (enriched for top-of-the-crypt cells) revealed that this subset, although transcriptionally heterogeneous, was almost exclusively composed of cells expressing high levels of genes characteristic of mature enterocytes (e.g., *CA1*$^{+}$, *CA2*$^{+}$, *KRT20*$^{+}$, *SLC26A3*$^{+}$, *AQP8*$^{+}$ and *MS4A12*$^{+}$)[14,21–23] and led to the discovery of at least two gene expression markers whose differential expression pattern—to our knowledge—has not been reported before (*CD177* and *GUCA2B*) (**Fig. 1g**). To validate the reliability of single-cell PCR gene-expression analysis results, we evaluated the distribution of SLC26A3 and CD177 protein expression in tissue sections and we confirmed its preferential expression at the top of the human colonic crypts (**Supplementary Figs. 11** and **12**).

We could also distinguish different subsets of cells with different transcriptional profiles within the EpCAM^{+}/CD44^{-}/CD66a^{high} population (e.g., *CA1*$^{+}$/*SLC26A3*$^{+}$ versus *GUCA2B*$^{+}$). At the present time, it is not clear whether they represent distinct stages of differentiation or distinct functional subsets of colonic enterocytes. Nonetheless, their clearly unique transcriptional programs identify them as part of a distinct cellular population.

Analysis of the EpCAM^{high}/CD44^{+} population (enriched for 'bottom-of-the-crypt' cells) revealed the presence of multiple populations, including: (i) a cell compartment characterized by the expression of genes linked to goblet cells (*MUC2*$^{+}$, *TFF3*^{high}, *SPDEF*$^{+}$, *SPINK4*$^{+}$)[24,25], (ii) a cell compartment characterized by the co-expression of genes associated with immature cells as well as genes known to be expressed by enterocytes (*OLFM4*$^{+}$, *CA2*^{high}) and (iii) a cell compartment whose gene-expression profile mirrors that of a stem and/or progenitor cell compartment in the mouse small intestine (*LGR5*$^{+}$, *ASCL2*$^{+}$, *PTPRO*$^{+}$, *RGMB*$^{+}$)[17,26]. A synopsis of the key genes that define the gene-expression profile of the different populations is provided in **Supplementary Table 3**.

The *OLMF4*$^{+}$/*CA2*^{high} and the *LGR5*$^{+}$/*ASCL2*$^{+}$ compartments shared expression of several genes of functional interest in both stem cell and cancer biology, such as genes involved in self-renewal and chromatin remodeling (*EZH2*, *BMI1*)[27–29], Wnt-pathway signaling (*AXIN2*)[30], cell growth and chemotaxis (*CXCL2*)[31], stem cell quiescence (*LRIG1*)[32] and oncogenes (*MYC*)[33]. The expression of proliferation markers, such as, *MKI67*, *TOP2A*, *BIRC5* (also known as *Survivin*) appeared to be restricted to the EpCAM^{high}/CD44^{+} (bottom-of-the-crypt) population and particularly to the *LGR5*$^{+}$/*ASCL2*$^{+}$ and *MUC2*$^{+}$/*TFF3*^{high} cells. This was partially expected based on both previously published data[14,17,19] and our own immunohistochemistry results (**Supplementary Fig. 13c**).

We also observed that *MUC2*$^{+}$/*TFF3*^{high} cells were characterized by high expression levels of several genes of interest, including *DLL1* and *DLL4*, encoding for two Notch ligands, and *KRT20*. The expression of *KRT20* at the bottom of the crypt appears contrary to the notion of KRT20 as a terminal differentiation marker. However, a more careful examination of immunohistochemical stainings identified scattered KRT20^{+} cells, which can be morphologically identified as goblet cells (**Supplementary Fig. 13a**,**b**). We also noticed that *MUC2*$^{+}$/*TFF3*^{high} cells, for the most part, did not express *CFTR*, the gene mutated in cystic fibrosis. The differential expression of *DLL4* is of potential relevance to the clinical development of novel anti-tumor therapeutic agents directed against this molecule[34].

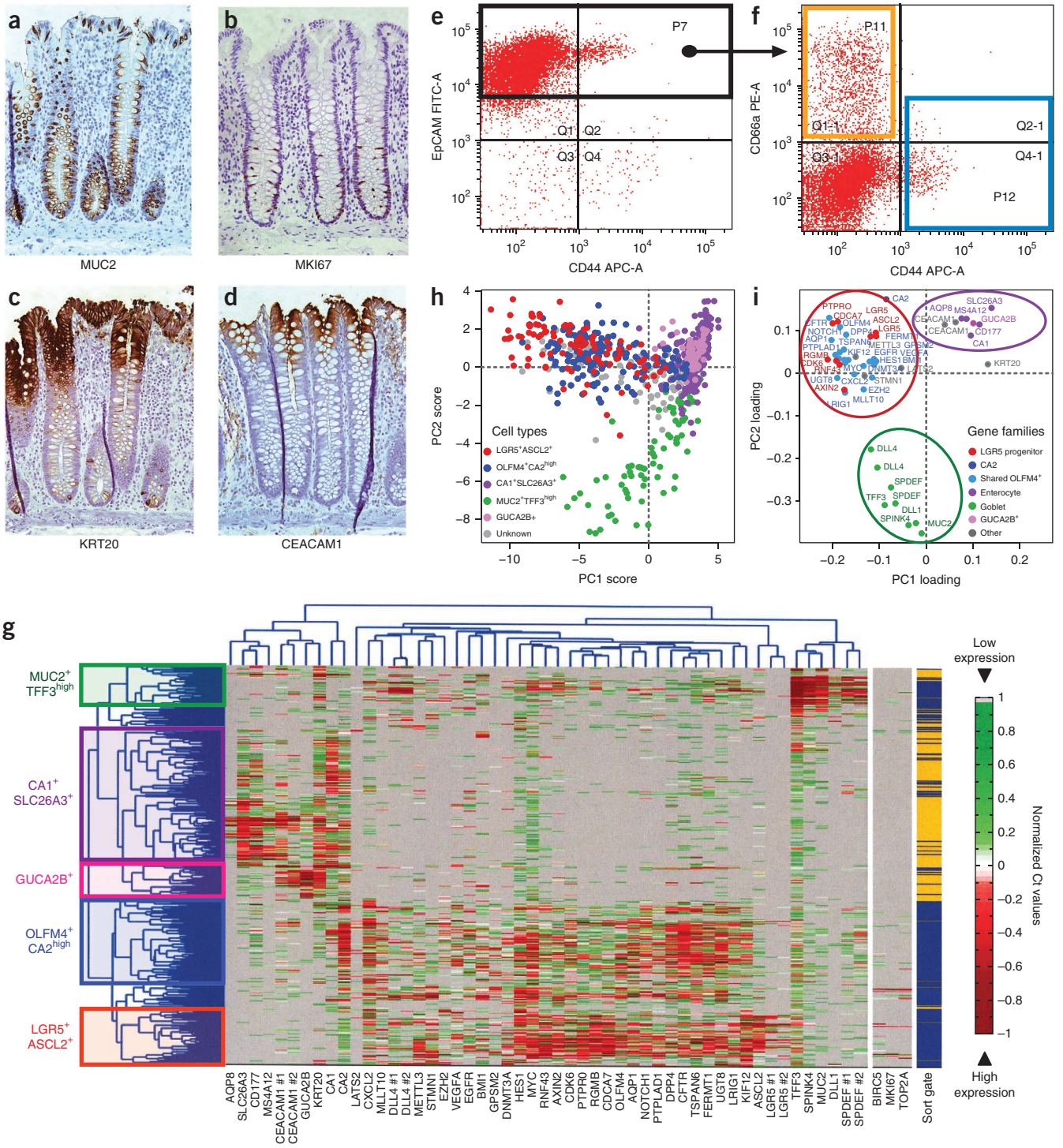


Figure 1 Single-cell PCR gene-expression analysis of human normal colon epithelium. (**a**–**d**) Immunohistochemistry of normal human colon epithelium, stained for MUC2 (**a**), labeling goblet cells, MKI67 (**b**), labeling proliferating cells, KRT20 (**c**) and CEACAM1 (**d**), preferentially labeling top-of-the-crypt cells. (**e**,**f**) Flow cytometry sorting strategy for top-of-the-crypt and bottom-of-the-crypt epithelial cells. (**e**) Colon epithelial cells, both $CD44^{neg}$ and $CD44^{+}$, were separated from stromal cells based on their $EpCAM^{+}$ phenotype. (**f**) Bottom-of-the-crypt epithelial cells were defined as $EpCAM^{high}/CD44^{+}$ (**f**, P12 blue sort gate) and top-of-the-crypt epithelial cells as $EpCAM^{+}/CD44^{-}/CD66a^{high}$ (**f**, P11 orange sort gate). (**g**) Hierarchical clustering of single-cell PCR gene-expression analysis data visualized distinct cell populations, including enterocyte-like cells (*CA1*$^{+}$/*SLC26A3*$^{+}$ and *GUCA2B*$^{+}$), goblet-like cells (*MUC2*$^{+}$/*TFF3*^{high}) and two compartments defined by gene-expression profiles reminiscent of more immature progenitors (*OLFM4*$^{+}$/*CA2*^{high} and *LGR5*$^{+}$/*ASCL2*$^{+}$). (**h**,**i**) Principal component analysis of single-cell PCR gene-expression data visualized different cell types and different gene families. Different cell types were characterized by different scores along the two main principal components (PC1 and PC2) (**h**). Different gene families were characterized by different contributions to the two main principal components. To allow comparisons between hierarchical clustering and PCA results, we displayed each cell or gene in PCA plots with the color corresponding to the cell type or gene family it was assigned to based on hierarchical clustering (**i**).

Analysis of a primary human colon adenoma

We then turned to cancer and investigated whether the cellular composition of the normal colonic epithelium is preserved in colorectal tumors, both benign and malignant. Analysis by single-cell PCR gene-expression analysis of EpCAM^{high}/CD44^{+} cells from a primary tubulo-villous adenoma (sample name: SU-COLON#76; **Supplementary Table 4**) revealed the presence of at least two different cell populations (that is, *LGR5*$^{+}$/*ASCL2*$^{+}$ and *MUC2*$^{+}$/*TFF3*^{high}) characterized by distinctive gene signatures, closely mirroring the subpopulations observed in corresponding EpCAM^{high}/CD44^{+} populations of normal tissues (**Fig. 2a–c**).

These observations were confirmed at the protein level by parallel immunohistochemical investigations for KRT20 and MUC2 (**Fig. 2d,e**) and are in agreement with the recent finding that

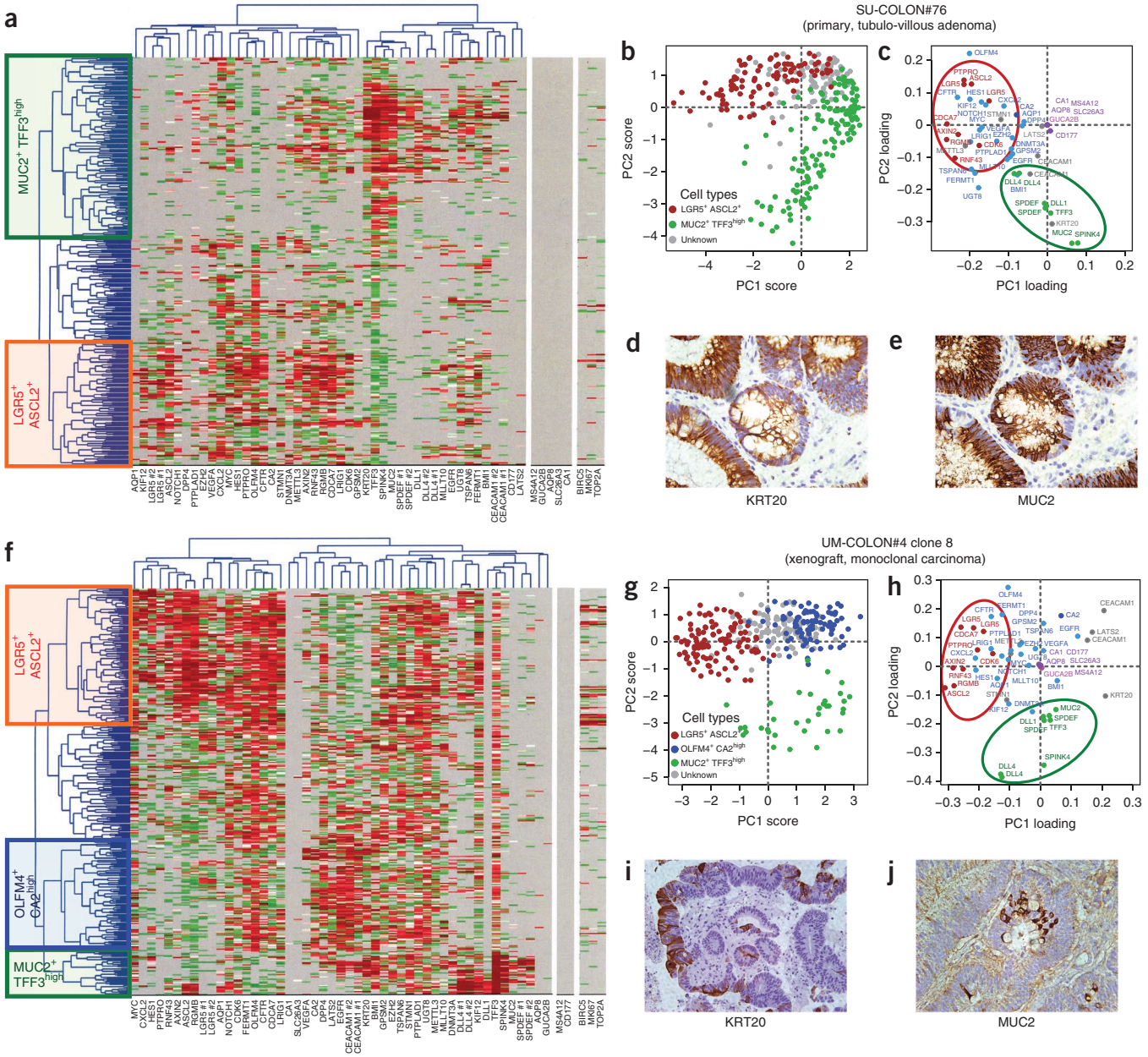


Figure 2 Single-cell PCR gene-expression analysis of human colon tumor tissues. (**a**) Hierarchical clustering of single-cell PCR gene-expression data from the EpCAM^{+}/CD44^{+} population of a large primary benign adenoma (sample: SU-COLON#76; see **Supplementary Table 4**). The analysis revealed the presence of multiple cell populations characterized by distinct gene signatures, closely mirroring lineages and differentiation stages observed in the EpCAM^{+}/CD44^{+} population from the normal colon epithelium. (**b**,**c**) Principal component analysis (PCA) of single-cell PCR gene-expression analysis data confirmed hierarchical clustering results, visualizing cell types (**b**) and gene families (**c**) similar to those identified in normal tissues. (**d**,**e**) Gene-expression data were confirmed at the protein level by immunohistochemistry, testing for expression of KRT20 (**d**) and MUC2 (**e**) on corresponding tissue sections. (**f**–**j**) A similar study on a monoclonal colon cancer xenograft obtained from injection of a single ($n = 1$) cell in a NOD/SCID/IL2Rγ$^{-/-}$ mouse (UM-COLON#4 clone 8) produced similar results in terms of hierarchical clustering (**f**), cell types identified by PCA (**g**), gene families identified by PCA (**h**), immunohistochemistry results for KRT20 (**i**) and immunohistochemistry results for MUC2 (**j**). Results from the monoclonal tumor xenograft indicated that the distinct cell populations visualized by single-cell PCR did not arise as the result of the coexistence within the tumor tissue of independent genetic subclones, but as the result of multilineage differentiation processes during tumor growth. Color coding of normalized threshold cycle (Ct) values in hierarchical clustering plots and of gene families in PC loading plots are identical to those of **Figure 1**.

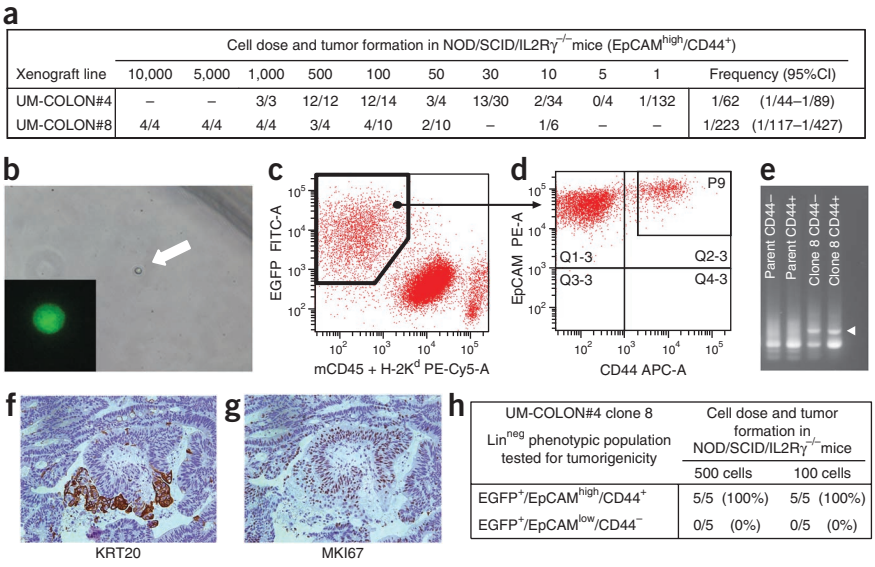


Figure 3 Analysis of a monoclonal human colon cancer xenograft obtained from injection of a single ($n = 1$) cell in NOD/SCID/IL2Rγ$^{-/-}$ mice. (**a**) In human colon cancer, the frequency of EpCAM^{high}/CD44^{+} cells capable to establish a tumor upon xenotransplantation in NOD/SCID/IL2Rγ$^{-/-}$ mice varies based on the xenograft line, as shown by comparative limiting-dilution experiments. (**b**) Single ($n = 1$) lentivirus-infected EGFP^{+}/EpCAM^{high}/CD44^{+} cancer cells can be sorted by flow cytometry for injection in mice. (**c**,**d**) Analysis by flow cytometry of a monoclonal tumor derived from injection of a single ($n = 1$), lentivirus-tagged, EGFP^{+}/EpCAM^{high}/CD44^{+} cancer cell from the human colon cancer xenograft UM-COLON#4 (clone 8) confirmed that human cells expressed EGFP (**c**) and contained both EpCAM^{low}/CD44^{-} and EpCAM^{high}/CD44^{+} populations (**d**). (**e**) The monoclonal origin of the UM-COLON#4 clone 8 tumor was confirmed by LM-PCR, showing the presence of a unique lentivirus integration site in both EGFP^{+}/EpCAM^{low}/CD44^{-} and EGFP^{+}/EpCAM^{high}/CD44^{+} populations, contrary to what was observed in its polyclonal parent tumor. A larger image of the LM-PCR gel is provided in **Supplementary Figure 24**. (**f**,**g**) Immunohistochemistry of monoclonal tumor tissues revealed heterogeneous and mutually exclusive expression patterns of KRT20 (**f**) and MKI67 (**g**). (**h**) Similar to what is observed in parent tumors, EpCAM^{high}/CD44^{+} and EpCAM^{low}/CD44^{-} populations from UM-COLON#4 clone 8 were characterized by different tumorigenic capacity, as evaluated by tumorigenicity experiments in NOD/SCID/IL2Rγ$^{-/-}$ mice.

KRT20 is frequently expressed in a mutually exclusive pattern with respect to *LGR5* (ref. 19). This primary adenoma appeared depleted in *CA1*$^{+}$/*SLC26A3*$^{+}$, *GUCA2B*$^{+}$and *OLFM4*$^{+}$/*CA2*^{high} cell populations. A careful examination of public gene-expression array databases indicated that this unexpected feature is likely common to many benign adenomas (**Supplementary Fig. 14**).

Analysis of a human colon cancer xenograft derived from a single cancer cell

Tumor tissues, both benign and malignant, are known to undergo perturbations of normal differentiation processes, but it is unclear to what extent those perturbations reflect quantitative changes in cell composition or qualitative changes in gene-expression programs. This topic has historically been controversial[4–9,35]. Our own systematic study of KRT20 and MUC2 protein expression in human malignant colorectal cancer tissues, for instance, revealed that both markers are frequently expressed heterogeneously, in patterns that mirror those observed in normal colorectal epithelium (**Supplementary Fig. 15**). It remained unclear, however, to what extent cancer transcriptional heterogeneity is the result of clonal genetic heterogeneity[36] or epigenetic heterogeneity due to multilineage differentiation processes[9].

To address this question from a functional perspective, we investigated whether a single ($n = 1$) human colorectal cancer cell can recreate the heterogeneous cell composition of parent tumor tissues, including the subpopulations that we discovered in this study. We injected NOD/SCID/IL2Rγ$^{-/-}$mice with single ($n = 1$) EpCAM^{high}/CD44^{+} cancer cells purified by flow cytometry from one of our well-characterized solid xenograft lines[37], following infection with a lentivirus vector encoding enhanced green fluorescence protein (EGFP; **Fig. 3a**,**b**).

Notably, the single cell–derived, lentivirus-tagged, EGFP^{+} xenograft line generated in this experiment (UM-COLON#4 clone 8) closely reproduced the phenotypic diversity of its parent tumor both in terms of tissue histology (**Figs. 2i**,**j** and **3f**,**g**) and surface-marker phenotypic repertoire of cellular populations (**Fig. 3c**,**d**). The line's monoclonal origin was confirmed by identification of a unique lentivirus integration site in all cancer cells (**Fig. 3e**).

Tumorigenicity experiments done in NOD/SCID/IL2Rγ$^{-/-}$ mice revealed that, as observed in the parent tumors[37], EGFP^{+}/EpCAM^{high}/CD44^{+} and EGFP^{+}/EpCAM^{low}/CD44$^{-/low}$ cell populations were endowed with different tumorigenic capacity (**Fig. 3h**). A single-cell PCR gene-expression analysis of the EpCAM^{high}/CD44^{+} population from these monoclonal tumors demonstrated its heterogeneous lineage composition, showing the presence of three distinct compartments (that is, *LGR5*$^{+}$/*ASCL2*$^{+}$, *OLFM4*$^{+}$/*CA2*^{high}, *MUC2*$^{+}$/*TFF3*^{high}), again characterized by distinctive gene signatures, closely mirroring those observed in corresponding immature populations of normal tissues (**Fig. 2f**–**h**).

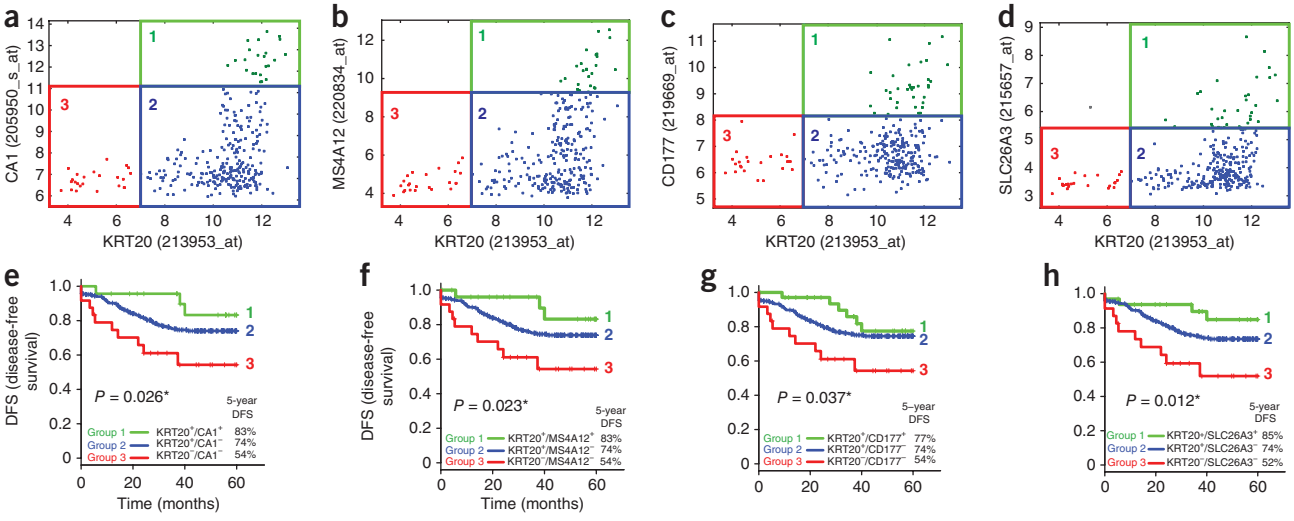
Taken together, these data formally prove that, in a subset of tumors, transcriptional heterogeneity is, at least partly, explained by multilineage differentiation processes that tend to recapitulate those observed in normal tissues.

Prognostic role of biomarkers identified by single-cell PCR

To gain further insight into the potential functional implications of these observations, we compared the gene-expression pattern of genes associated with cell proliferation (that is, *MKI67*, *TOP2A* and *BIRC5*) in normal and cancer tissues. In this case too, we observed that the expression pattern observed in malignant tissues frequently mirrored that of normal ones.

Both in the normal tissue and in the monoclonal human colon cancer xenograft, for instance, all three proliferation markers were frequently expressed in a mutually exclusive way as compared to the differentiation marker *KRT20* (**Supplementary Fig. 16**). This observation was subsequently confirmed at the protein level by a systematic study of MKI67 and KRT20 expression in serial sections from seven human colorectal cancer tissues, where MKI67 expression was often inversely associated with KRT20 (**Supplementary Fig. 17**).

These observations suggest that, in at least some cases, bulk short-term tumor growth is principally driven by a specific subset of the cancer cell population, characterized by a gene-expression repertoire characteristic of more immature cell compartments. This concept has important implications for the modeling of tumor growth kinetics



i KRT20/CA1

Variable	HR	95% CI	*P* value
Group (1–3)	2.44	1.37–4.35	0.0024*
Stage (I–IV)	3.40	2.42–4.77	< 0.001**
Age	0.99	0.97–1.01	0.19
Sex (M vs. F)	1.11	0.87–1.42	0.39

j KRT20/MS4A12

Variable	HR	95% CI	*P* value
Group (1–3)	2.43	1.38–4.29	0.0022*
Stage (I–IV)	3.38	2.41–4.75	< 0.001**
Age	0.99	0.97–1.01	0.17
Sex (M vs. F)	1.14	0.89–1.46	0.28

k KRT20/CD177

Variable	HR	95% CI	*P* value
Group (1–3)	2.16	1.27–3.68	0.0045*
Stage (I–IV)	3.34	2.39–4.66	< 0.001**
Age	0.99	0.97–1.00	0.14
Sex (M vs. F)	1.13	0.88–1.44	0.33

l KRT20/SLC26A3

Variable	HR	95% CI	*P* value
Group (1–3)	2.38	1.36–4.18	0.0025*
Stage (I–IV)	3.35	2.39–4.69	< 0.001**
Age	0.99	0.97–1.01	0.19
Sex (M vs. F)	1.12	0.88–1.43	0.35

Figure 4 *KRT20* and *top-crypt* genes can be used as prognostic markers in colorectal cancer patients. (**a**–**d**) We used the Hegemon software to graph individual arrays according to the expression levels of *KRT20* and one of four genes characteristic of top-of-the-crypt *CA1*$^{+}$/*SLC26A3*$^{+}$ enterocyte-like cells: *KRT20* versus *CA1* (**a**), *KRT20* versus *MS4A12* (**b**), *KRT20* versus *CD177* (**c**), *KRT20* versus *SLC26A3* (**d**). We used the StepMiner algorithm to define gene-expression thresholds and identify three distinct gene-expression groups: Group 1 (green), defined as *KRT20*$^{+}$/*CA1*^{high}, *KRT20*$^{+}$/*MS4A12*^{high}, *KRT20*$^{+}$/*CD177*$^{+}$ or *KRT20*$^{+}$/*SLC26A3*$^{+}$, respectively; Group 2 (blue), defined as *KRT20*$^{+}$/*CA1*$^{-/low}$, *KRT20*$^{+}$/*MS4A12*$^{-/low}$, *KRT20*$^{+}$/*CD177*$^{-}$ or *KRT20*$^{+}$/*SLC26A3*$^{-}$, respectively; Group 3 (red), defined as *KRT20*$^{-}$/*CA1*$^{-/low}$, *KRT20*$^{-}$/*MS4A12*$^{-/low}$, *KRT20*$^{-}$/*CD177*$^{-}$ or *KRT20*$^{-}$/*SLC26A3*$^{-}$, respectively. (**e**–**h**) Survival analysis using Kaplan-Meier curves showed that, in all four cases, an increasingly immature gene-expression profile corresponded to a progressively worse prognosis. (**i**–**l**) Multivariate analysis of survival data based on the Cox proportional hazards model indicated that the prognostic effect of these two-gene classifiers was not confounded by clinical stage, age or sex. The analysis was performed on a pooled database of 299 primary colon cancer gene-expression arrays annotated with disease-free survival (DFS) data[41,42] (**Supplementary Table 1**). *$P < 0.05$, **$P < 0.001$. Age modeled as a continuous variable. HR, hazard ratio; CI, confidence interval; M, male; F, female.

and the response to anti-tumor drugs in different experimental settings. Although very common, this feature is not absolute, as we have observed exceptions characterized either by homogenous expression of KRT20 in almost the entirety of the malignant epithelium or by complete absence of it in selected human tumors (**Supplementary Fig. 17**, samples SU87 and SU98, respectively). In accordance with our model, tumors characterized by the complete absence of KRT20 expression were very poorly differentiated and contained high percentages of MKI67^{+} cells (**Supplementary Fig. 17**, SU98).

We next tested whether these insights in the functional anatomy of the colon epithelium could have clinically useful applications. We evaluated whether quantitative expression levels of genes associated with differentiation processes could be used as a substitute measure for the cellular composition of the corresponding tumors and thereby serve to stratify colon cancer patients and predict clinical outcome. Our single-cell PCR gene-expression analysis data identified a set of sensitive and exclusive markers of top-of-the-crypt *CA1*$^{+}$/*SLC26A3*$^{+}$ cells (that is, *CA1*, *MS4A12*, *CD177*, *SLC26A3*). It also implicated KRT20 as a more promiscuous differentiation marker, whose expression is high in *CA1*$^{+}$/*SLC26A3*$^{+}$ cells and a subset of *MUC2*$^{+}$/*TFF3*^{high} cells, is absent in *LGR5*$^{+}$/*ASCL2*$^{+}$ cells, and is inversely associated with that of proliferation markers (*MKI67*, *TOP2A*, *BIRC5*). In addition, KRT20 expression can be easily detected by immunohistochemistry and is commonly used as a diagnostic marker in surgical pathology[38], thus representing an attractive candidate for further clinical applications[39].

Our first analysis of a pool of 1,568 independent human colon gene-expression arrays revealed that expression levels of genes characteristic for the *CA1*$^{+}$/*SLC26A3*$^{+}$ cell population are strongly correlated (**Supplementary Fig. 18**). The relationship between the expression of these top-of-the-crypt genes and *KRT20* was described by a Boolean implication: tumors expressing high levels of top-of-the-crypt genes (*top-crypt*^{high}) were always *KRT20*$^{+}$, whereas tumors expressing low-to-negative levels of top-of-the-crypt genes (top-crypt$^{-/low}$) could be clearly separated into two groups: *KRT20*$^{+}$ and *KRT20*$^{-}$ (**Supplementary Fig. 7**). Importantly, *KRT20*$^{-}$ tumors expressed high levels of *ALCAM/CD166* (**Supplementary Fig. 19**), a gene encoding for a surface marker characteristic of colon cancer cells with high tumorigenic potential in mouse xenotransplantation experiments[37].

We developed software ('hierarchical exploration of gene expression microarrays on-line', or Hegemon) to analyze the survival outcomes of human colon cancer patients after stratification into distinct gene-expression subsets, based on the expression of *KRT20* and one of the marker genes of *CA1*$^{+}$/*SLC26A3*$^{+}$ top-of-the-crypt cells (**Fig. 4a**–**d**). These subsets, or gene-expression groups, were numbered from more to less mature (group 1, *KRT20*$^{+}$/*top-crypt*^{high}; group 2, *KRT20*$^{+}$/*top-crypt*$^{-/low}$; group 3, *KRT20*$^{-}$/*top-crypt*$^{-/low}$). We used a computer-assisted method to determine the threshold level

Table 1 The prognostic effect of *KRT20/top-crypt* gene-expression groups

	HR[a]	95% CI[b]	*P* value
KRT20/CA1			
Prognostic variable			
Group (1–3) KRT20/CA1	2.93	1.37–6.27	0.0056*
Grade (G1–G4)	1.09	0.58–2.04	0.80
Stage (I–IV)	3.43	2.20–5.34	< 0.001**
Age[c]	0.99	0.97–1.01	0.43
Sex (M/F)[d]	1.18	0.86–1.61	0.31
KRT20/MS4A12			
Prognostic variable			
Group (1–3) KRT20/MS4A12	2.93	1.37–6.28	0.0057*
Grade (G1–G4)	1.07	0.57–2.00	0.84
Stage (I–IV)	3.41	2.19–5.31	<0.001**
Age[c]	0.99	0.97–1.01	0.41
Sex (M/F)[d]	1.19	0.87–1.63	0.28
KRT20/CD177			
Prognostic variable			
Group (1–3) KRT20/CD177	1.94	0.97–3.90	0.062
Grade (G1–G4)	1.19	0.63–2.22	0.59
Stage (I–IV)	3.21	3.03–7.06	<0.001**
Age[c]	0.99	0.97–1.01	0.39
Sex (M/F)[d]	1.20	0.87–1.64	0.26
KRT20/SLC26A3			
Prognostic variable			
Group (1–3) KRT20/SLC26A3	2.36	1.14–4.88	0.021*
Grade (G1–G4)	1.12	0.60–2.10	0.72
Stage (I–IV)	3.34	2.16–5.15	<0.001**
Age[c]	0.99	0.97–1.01	0.45
Sex (M/F)[d]	1.19	0.87–1.63	0.27

Multivariate analysis based on the Cox proportional hazards model, testing the *KRT20/top-crypt* two-gene scoring systems in parallel with pathological grading, clinical stage, age and sex, using the *KRT20/CA1* two-gene classifier, the *KRT20/MS4A12* two-gene classifier, the *KRT20/CD177* two-gene classifier or the *KRT20/ SLC26A3* two-gene classifier. Contrary to pathological grade, *KRT20/top-crypt* gene expression groups were associated with statistically significant ($p < 0.05$) hazard ratios (HR), with the only exception of the *KRT20/CD177* two-gene classifier. The analysis was performed on a subset database of 181 microarrays annotated with grading information (database from ref. 42, $n = 181$, see **Supplementary Table 1**). *, $P < 0.05$; **, $P < 0.001$. [a]HR, hazard-ratio. [b]CI, confidence interval. [c]Age modeled as a continuous variable. [d]M/F, male versus female.

between positive and negative expression, based on the StepMiner algorithm (**Supplementary Fig. 20**)[40], and compared the clinical outcome of colon cancer patients in the three groups, using a pool of three independent data sets, containing 299 patients at different clinical stages (either AJCC stage I–IV or Dukes stage A–D) from the H. Lee Moffit Cancer Center, the Vanderbilt Medical Center and the Royal Melbourne Hospital[41,42], all of which were annotated with disease-free survival (DFS) data.

The three patient groups identified by these simple two-gene classifiers displayed substantially different clinical outcomes. An increasingly immature gene-expression profile corresponded to a progressively worse prognosis (**Fig. 4e–h**). This result was independent of the gene chosen as marker of $CA1^{+}/SLC26A3^{+}$ cells (that is, *CA1*, *MS4A12*, *CD177*, *SLC26A3*) and a multivariate analysis indicated that the prognostic value of the two-gene grouping system was not confounded by stage or other clinical variables (**Fig. 4i–l**).

Tumors with a more immature gene-expression profile (group 3, $KRT20^{-}/top\text{-}crypt^{-/low}$) were more likely to be of high pathological grade (G3–G4; **Supplementary Fig. 21**) and of microsatellite instability status (MSI; **Supplementary Fig. 22**). These enrichments, however, did not confound the prognostic value of the two-gene classifier system, as the high hazard-ratios associated with more immature gene-expression groups remained statistically significant ($P < 0.05$), when tested against pathological grade in multivariate analysis (**Table 1**; with the exception of *KRT20/CD177*, $P = 0.06$), and because MSI^{+} tumors are known to be usually associated with a better prognosis[43]. The prognostic effect of the two-gene classifier system was also independent of the recently described multigene EphB2 intestinal stem cell signature[19], and was associated with comparable, if not superior, hazard ratios (**Supplementary Fig. 23**).

DISCUSSION

In this study, we implemented a method to investigate the cellular composition of solid tissues based on high-throughput parallel analysis of the gene-expression repertoire of single cells sorted by flow cytometry. We used this methodology to identify distinct cellular subsets of the human colon epithelium and to discover gene expression markers to define them. We then examined human colorectal tumors, both benign and malignant, and characterized them in terms of cell lineage composition and maturation. We showed that tumor tissues contain multiple cell types whose transcriptional identities mirror those of the cellular lineages of the normal epithelium. Moreover, we showed that tumor tissues generated from a single cell can recapitulate the lineage diversity of parent tumors, demonstrating that multilineage differentiation represents a key source of *in vivo* functional and phenotypic cancer cell heterogeneity.

Using these concepts as a guide, we identified biological subsets of human colorectal cancer, based on the expression of genes characteristic of specific cell types. These biological subsets were associated with substantially different clinical outcomes and could be identified by a simple two-gene classifier system. This prognostic scoring system appeared independent of and superior to pathological grading, which is, to this date, one of the few parameters incorporated into the design of therapeutic algorithms for colon cancer patients[44]. Owing to its simplicity and quantitative nature, this two-gene scoring system has the potential to move beyond the realm of purely experimental medicine and is a viable candidate for clinical applications.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This study was supported by National Institutes of Health (NIH) grants U54-CA126524 and P01-CA139490 (to S.R.Q. and M.F.C.), the NIH Director's Pioneer Awards (to S.R.Q.) and a grant from the Ludwig foundation (to M.F.C.). P.D. was supported by a training grant from the California Institute for Regenerative Medicine (CIRM) and by a BD Biosciences Stem Cell Research Grant (Summer 2011). T.K. was supported by a fellowship from the Machiah Foundation. D.S. was supported by NIH grant K99-CA151673, by Department of Defense grant W81XWH-10-1-0500 and a grant from the Siebel Stem Cell Institute and the Thomas and Stacey Siebel Foundation. We wish to thank R. Tibshirani, D. Witten, L. Warren, R.A. White III, E. Gilbert, P. Lovelace, M. Palmor, C. Donkers and S.P. Miranda for helpful discussion and technical support in many moments during the completion of this study.

AUTHOR CONTRIBUTIONS

P.D., T.K., D.S., M.F.C. and S.R.Q. conceived the study and designed the experiments. P.S.R., M.E.R., A.A.L., M.Z., N.F.N., M.v.d.W. and H.C. provided intellectual guidance in the design of selected experiments. P.D., T.K., D.S., P.S.R., A.A.L., S.S., J.O., D.M.J., D.Q., J.W., and S.H. performed the experiments. P.D., T.K., D.S., N.F.N., Y.S., M.F.C. and S.R.Q. analyzed the data and/or provided intellectual guidance in their interpretation. J.B., A.A.S. and B.V. provided samples and reagents. P.D., T.K., D.S., M.F.C. and S.R.Q. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/nbt/index.html.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Reya, T., Morrison, S.J., Clarke, M.F. & Weissman, I.L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
2. Jordan, C.T., Guzman, M.L. & Noble, M. Cancer stem cells. *N. Engl. J. Med.* **355**, 1253–1261 (2006).
3. Dalerba, P., Cho, R.W. & Clarke, M.F. Cancer stem cells: models and concepts. *Annu. Rev. Med.* **58**, 267–284 (2007).
4. Shackleton, M., Quintana, E., Fearon, E.R. & Morrison, S.J. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell* **138**, 822–829 (2009).
5. Campbell, L.L. & Polyak, K. Breast tumor heterogeneity: cancer stem cells or clonal evolution? *Cell Cycle* **6**, 2332–2338 (2007).
6. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805**, 105–117 (2010).
7. Kirkland, S.C. Clonal origin of columnar, mucous, and endocrine cell lineages in human colorectal epithelium. *Cancer* **61**, 1359–1363 (1988).
8. Odoux, C. *et al.* A stochastic model for cancer stem cell origin in metastatic colon cancer. *Cancer Res.* **68**, 6932–6941 (2008).
9. Vermeulen, L. *et al.* Single-cell cloning of colon cancer stem cells reveals a multi-lineage differentiation capacity. *Proc. Natl. Acad. Sci. USA* **105**, 13427–13432 (2008).
10. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469**, 415–418 (2011).
11. Warren, L., Bryder, D., Weissman, I.L. & Quake, S.R. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc. Natl. Acad. Sci. USA* **103**, 17807–17812 (2006).
12. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene-expression analysis from zygote to blastocyst. *Dev. Cell* **18**, 675–685 (2010).
13. White, A.K. *et al.* High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl. Acad. Sci. USA* **108**, 13999–14004 (2011).
14. Jiao, Y.F., Nakamura, S., Sugai, T., Yamada, N. & Habano, W. Serrated adenoma of the colorectum undergoes a proliferation versus differentiation process: new conceptual interpretation of morphogenesis. *Oncology* **74**, 127–134 (2008).
15. Wielenga, V.J. *et al.* Expression of CD44 in Apc and Tcf mutant mice implies regulation by the WNT pathway. *Am. J. Pathol.* **154**, 515–523 (1999).
16. Prall, F. *et al.* CD66a (BGP), an adhesion molecule of the carcinoembryonic antigen family, is expressed in epithelium, endothelium, and myeloid cells in a wide range of normal human tissues. *J. Histochem. Cytochem.* **44**, 35–41 (1996).
17. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003–1007 (2007).
18. Becker, L., Huang, Q. & Mashimo, H. Immunostaining of Lgr5, an intestinal stem cell marker, in normal and premalignant human gastrointestinal tissue. *Scientific World Journal* **8**, 1168–1176 (2008).
19. Merlos-Suarez, A. *et al.* The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8**, 511–524 (2011).
20. Sahoo, D., Dill, D.L., Gentles, A.J., Tibshirani, R. & Plevritis, S.K. Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol.* **9**, R157 (2008).
21. Hoglund, P. *et al.* Mutations of the down-regulated in adenoma (DRA) gene cause congenital chloride diarrhoea. *Nat. Genet.* **14**, 316–319 (1996).
22. Fischer, H., Stenling, R., Rubio, C. & Lindblom, A. Differential expression of aquaporin 8 in human colonic epithelial cells and colorectal tumors. *BMC Physiol.* **1**, 1 (2001).
23. Koslowski, M., Sahin, U., Dhaene, K., Huber, C. & Tureci, O. MS4A12 is a colon-selective store-operated calcium channel promoting malignant cell processes. *Cancer Res.* **68**, 3458–3466 (2008).
24. Noah, T.K., Kazanjian, A., Whitsett, J. & Shroyer, N.F. SAM pointed domain ETS factor (SPDEF) regulates terminal differentiation and maturation of intestinal goblet cells. *Exp. Cell Res.* **316**, 452–465 (2010).
25. Gregorieff, A. *et al.* The ets-domain transcription factor Spdef promotes maturation of goblet and paneth cells in the intestinal epithelium. *Gastroenterology* **137**, 1333–1345 (2009).
26. van der Flier, L.G. *et al.* Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* **136**, 903–912 (2009).
27. Ezhkova, E. *et al.* Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. *Cell* **136**, 1122–1135 (2009).
28. Park, I.K. *et al.* Bmi-1 is required for maintenance of adult self-renewing haematopoietic stem cells. *Nature* **423**, 302–305 (2003).
29. Sangiorgi, E. & Capecchi, M.R. Bmi1 is expressed *in vivo* in intestinal stem cells. *Nat. Genet.* **40**, 915–920 (2008).
30. Zeng, Y.A. & Nusse, R. Wnt proteins are self-renewal factors for mammary stem cells and promote their long-term expansion in culture. *Cell Stem Cell* **6**, 568–577 (2010).
31. Beider, K., Abraham, M. & Peled, A. Chemokines and chemokine receptors in stem cell circulation. *Front. Biosci.* **13**, 6820–6833 (2008).
32. Jensen, K.B. *et al.* Lrig1 expression defines a distinct multipotent stem cell population in mammalian epidermis. *Cell Stem Cell* **4**, 427–439 (2009).
33. Dalla-Favera, R., Wong-Staal, F. & Gallo, R.C. Onc gene amplification in promyelocytic leukaemia cell line HL-60 and primary leukaemic cells of the same patient. *Nature* **299**, 61–63 (1982).
34. Hoey, T. *et al.* DLL4 blockade inhibits tumor growth and reduces tumor-initiating cell frequency. *Cell Stem Cell* **5**, 168–177 (2009).
35. Park, S.Y., Gonen, M., Kim, H.J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120**, 636–644 (2010).
36. Losi, L., Baisse, B., Bouzourene, H. & Benhattar, J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. *Carcinogenesis* **26**, 916–922 (2005).
37. Dalerba, P. *et al.* Phenotypic characterization of human colorectal cancer stem cells. *Proc. Natl. Acad. Sci. USA* **104**, 10158–10163 (2007).
38. Oien, K.A. Pathologic evaluation of unknown primary cancer. *Semin. Oncol.* **36**, 8–37 (2009).
39. Lugli, A., Tzankov, A., Zlobec, I. & Terracciano, L.M. Differential diagnostic and functional role of the multi-marker phenotype CDX2/CK20/CK7 in colorectal cancer stratified by mismatch repair status. *Mod. Pathol.* **21**, 1403–1412 (2008).
40. Sahoo, D., Dill, D.L., Tibshirani, R. & Plevritis, S.K. Extracting binary signals from microarray time-course data. *Nucleic Acids Res.* **35**, 3705–3712 (2007).
41. Jorissen, R.N. *et al.* Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer. *Clin. Cancer Res.* **15**, 7642–7651 (2009).
42. Smith, J.J. *et al.* Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology* **138**, 958–968 (2010).
43. Guastadisegni, C., Colafranceschi, M., Ottini, L. & Dogliotti, E. Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. *Eur. J. Cancer* **46**, 2788–2798 (2010).
44. Bardia, A. *et al.* Adjuvant chemotherapy for resected stage II and III colon cancer: comparison of two widely used prognostic calculators. *Semin. Oncol.* **37**, 39–46 (2010).

ONLINE METHODS

Human primary tissues and colon cancer xenografts. Human primary colon tissues, normal or malignant, were collected according to guidelines from Stanford University's institutional review board. Human colon cancer tissues used in this study, either from primary samples or xenograft lines, are listed in **Supplementary Table 4**, together with clinical information relative to corresponding patients. Human colon cancer xenograft lines were established and serially passaged in immunodeficient mice following previously published protocols[37]. A detailed description of these protocols is provided in the **Supplementary Methods**.

Cell lines. Calibration experiments to measure accuracy and precision of single-cell sorting by flow cytometry, as well as to measure single-cell sensitivity of single-cell PCR, were performed on a clone of the HCT116 human colon cancer cell line infected with the pLL3.7 lentivirus (Addgene no. 11795). HCT116 cells are available from the American Tissue-type Culture Collection (ATCC; CCL-247) and were maintained in RPMI-1640 medium, supplemented with 10% heat-inactivated fetal bovine serum, 2 mM L-glutamine, 120 μg/ml penicillin, 100 μg/ml streptomycin, 20 mM HEPES and 1 mM sodium pyruvate, as previously described[45].

Solid tissue disaggregation. Solid tissues, normal and neoplastic, collected from primary surgical specimens or mouse xenografts, were mechanically and enzymatically disaggregated into single-cell suspensions, following previously published protocols[37]. Briefly, solid tissues were minced into small chunks (2 mm^3), rinsed with Hank's balanced salt solution (HBSS), finely chopped with a razor blade into minute fragments (0.2–0.5 mm^3), resuspended in serum-free RPMI-1640 medium (2 mM L-glutamine, 120 μg/ml penicillin, 100 μg/ml streptomycin, 50 μg/ml ceftazidime, 0.25 μg/ml amphotericin-B, 20 mM HEPES, 1 mM sodium pyruvate) together with 100 units/ml DNase-I and 200 units/ml Collagenase-III (Worthington) and incubated for 2 h at 37 °C to obtain enzymatic disaggregation. Cell suspensions were serially filtered with sterile gauze, 70-μm and 40-μm nylon meshes. Red blood cells were removed by osmotic lysis with ACK hypotonic buffer (150 mM NH_4Cl, 1 mM $KHCO_3$; 5 min on ice).

Flow cytometry and single-cell sorting experiments. To minimize loss of cell viability, we performed experiments on fresh cell suspensions, prepared shortly before flow cytometry[37]. Antibody staining was performed in HBSS supplemented with 2% heat-inactivated calf serum, 120 μg/ml penicillin, 100 μg/ml streptomycin, 50 μg/ml ceftazidime, 0.25 μg/ml amphotericin-B, 20 mM HEPES, 1 mM sodium pyruvate and 5 mM EDTA. To minimize unspecific binding of antibodies, cells were first incubated with 0.6% human IgGs (Gammagard Liquid; Baxter) for 10 min on ice, at a concentration of $3–5 \times 10^5$ cells/100 μl. Cells were subsequently washed and stained with antibodies at dilutions determined by appropriate titration experiments. Antibodies used in this study include anti-human EpCAM-FITC or PE (clone EBA-1; BD Biosciences), anti-human CD44-APC (clone G44–26; BD Biosciences) and anti-human CD66a-PE (clone 283340; R&D Systems). Cells positive for expression of nonepithelial lineage markers (Lin^+) were excluded by staining with PE.Cy5-labeled antibodies using different strategies for primary tissues and mouse xenografts. In experiments on primary human tissues, stromal cells were excluded by staining with anti-human CD3-biotin (clone UCHT1; BD Biosciences), CD16-biotin (clone 3G8; BD Biosciences), CD45-biotin (clone HI30; BD Biosciences), and CD64-biotin (clone 10.1; BD Biosciences) + streptavidin-PE/Cy5 (BD Biosciences). In experiments on human colon cancer xenografts, mouse cells were excluded by staining with anti-mouse CD45-PE/Cy5 (clone 30-F11; BD Biosciences) and anti-mouse H-$2K^d$-biotin (clone SF1–1.1; BD Biosciences) + streptavidin-PE/Cy5 (BD Biosciences). After 15 min on ice, stained cells were washed of excess unbound antibodies and resuspended in HBSS with 2% heat-inactivated calf serum, 20 mM HEPES, 5 mM EDTA, 1 mM sodium pyruvate and 1.1 μM DAPI dilactate (Molecular Probes). Flow-cytometry analysis was performed using a BD FACSAriaII cell-sorter (Becton Dickinson). Forward-scatter height versus forward-scatter width (FSC-H versus FSC-W) and side-scatter height versus side-scatter width (SSC-H versus SSC-W) profiles were used to eliminate cell doublets. Dead cells were eliminated by excluding $DAPI^+$ cells, whereas contaminating human or mouse Lin^+ cells were eliminated by excluding $PE/Cy5^+$ cells.

In single-cell sorting experiments, each single ($n = 1$) cell was individually sorted into a different well of a 96-well PCR plate, using a protocol already built-in within the FACSAriaII software package, with appropriate adjustments (device: 96-well plate; precision: single-cell; nozzle: 130 μm).

Single-cell PCR. Single-cell gene-expression experiments were performed using Fluidigm's M96 quantitative PCR (qPCR) DynamicArray microfluidic chips (Fluidigm). Single cells were sorted by FACS into individual wells of 96-well PCR plates as described above. Each 96-well plate was preloaded with 5 μl/well of CellsDirect PCR mix (Invitrogen) and 0.1 μl/well (2 U) of SuperaseIn RNase-inhibitor. Following single-cell sorting, each well was supplemented with 1 μl (Applied Biosystems) of SuperScript-III RT/Platinum Taq (Invitrogen), 1.5 μl of Tris-EDTA (TE) buffer and 2.5 μl of a mixture of 96 pooled TaqMan assays (Applied Biosystems) containing each assay at 1:100 dilution. Single-cell mRNA was directly reverse transcribed into cDNA (50 °C for 15 min, 95 °C for 2 min), pre-amplified for 20 cycles (each cycle: 95 °C for 15 s, 60 °C for 4 min) and diluted 1:3 with TE buffer. A 2.25 μl aliquot of amplified cDNA was then mixed with 2.5 μl of TaqMan Universal PCR Master Mix (Applied Biosystems) and 0.25 μl of Fluidigm's "sample loading agent," then inserted into one of the chip "sample" inlets. Individual TaqMan assays were diluted at 1:1 ratios with TE. A 2.5 μl aliquot of each diluted TaqMan assay was mixed with 2.5 μl of Fluidigm's "assay loading agent" and individually inserted into one of the chip "assay" inlets. Samples and probes were loaded into M96 chips using an IFC Controller HX (Fluidigm), then transferred to a BioMark real-time PCR reader (Fluidigm) following manufacturer's instructions. A list of the 57 TaqMan assays used in this study is provided in **Supplementary Table 2**.

Analysis and graphic display of single-cell PCR data. Single-cell PCR data were analyzed and displayed using MATLAB (MathWorks) as summarized in **Supplementary Figure 2**. A minimum of 336 cells were analyzed for each phenotypic population, corresponding to four PCR plates, each containing 84 single cells ($84 \times 4 = 336$), eight positive and four negative controls. As positive controls, we used replicates of a 1:1:1 mixture of total RNA standards from human normal colon (AM7986), human normal testes (AM7972) and HeLa cells (AM7852), all from Applied Biosystems. Results from cells not expressing *ACTB* (β-actin) and *GAPDH* (glyceraldehyde 3-phosphate dehydrogenase), or expressing them at extremely low values (Ct >35), were removed from the analysis. Gene-expression results were normalized by mean centering and dividing by 3 times the standard deviation (3 s.d.) of expressing cells (**Supplementary Fig. 2**), and visualized using both hierarchical clustering and PCA[12,46]. Hierarchical clustering was performed both on cells and genes, based on Euclidean or correlation distance metric and complete linkage. Positive or negative associations between two genes were tested by Spearman correlation, and *P*-values calculated based on 10,000 permutations. Both hierarchical clustering and PCA were based on the results for 47 differentially expressed genes (51 assays), and excluded results from housekeeping (*ACTB*, *GAPDH*, *EpCAM*) and proliferation-related genes (*MKI67*, *TOP2A*, *BIRC*) to avoid noise based on proliferation status. A detailed description of all these procedures is provided in the **Supplementary Methods**.

Immunohistochemistry and immunofluorescence. Paraffin-embedded tissue sections were stained with anti-human CK20 (clone Ks20.8, DakoCytomation), MUC2 (clone Ccp58, Fitzgerald Industries), MKI67 (clone MIB-1, DakoCytomation), CEACAM1/CD66a (clone 283340; R&D Systems) and SLC26A3 (lot no. R32905, Sigma Life Science) antibodies, according to manufacturers' instructions. Frozen tissue sections were stained with an anti-human CD177 antibody (clone MEM-166, BD Biosciences) followed by secondary staining with goat anti-mouse IgG-Alexa488 (Invitrogen). A description of immunohistochemistry and immunofluorescence protocols is provided in the **Supplementary Methods**.

Generation and characterization of monoclonal tumors. $EpCAM^{high}/CD44^+$ human colon cancer cells were infected with the pLL3.7 lentivirus (Addgene #11795)[47]. Cells were infected by spin-inoculation for 4 h and injected in bulk

doi:10.1038/nbt.2038

into the subcutaneous tissue of a NOD/SCID/IL2Rγ$^{-/-}$ mice. The resulting tumors were analyzed to evaluate infection efficiency, and EGFP^{+}/EpCAM^{high}/CD44^{+} were re-sorted and injected as single cells, again into NOD/SCID/IL2Rγ$^{-/-}$ mice. Monoclonal origin of tumors originated from single (n = 1) lentivirus-infected EpCAM^{high}/CD44^{+} cancer cells was confirmed by ligation-mediated PCR (LM-PCR)[48], followed by DNA sequencing of LM-PCR amplification products. In the case of UM-COLON#4 clone 8, DNA sequencing of LM-PCR amplification products pinpointed the provirus integration-site on the long arm of human chromosome 19 (19q13.3), in proximity of the *AP3D1* gene (adaptor-related protein complex 3, delta 1 subunit). For a visual guide on how to interpret LM-PCR results refer to **Supplementary Figure 24**.

Tumorigenicity experiments. Tumorigenicity experiments were performed in NOD/SCID/IL2Rγ$^{-/-}$ immunodeficient mice following previously published protocols[37,49,50] and Stanford University's institutional animal welfare guidelines. Tumorigenic cell frequencies were calculated by limiting dilution using the L-Calc software (StemCell Technologies). A detailed description of the protocols used for tumorigenicity experiments is provided in the **Supplementary Methods**.

Bioinformatic data collection and assemblage of the "human colon global database." A collection of 46,047 publicly available human gene-expression arrays (25,721 arrays on Affymetrix U133 Plus 2.0, 16,357 arrays on Affymetrix U133A, 3,969 arrays on Affymetrix U133A 2.0) was downloaded from NCBI's GEO database and normalized using the RMA (Robust Multi-chip Average) algorithm. Normalization was done either independently for each platform or on the whole array collection, using a modified CDF (chip description file) reduced to contain only shared probes. From this general collection, which contained arrays from all types of human samples, we extracted a subset of 1,684 unique arrays from human colon tissues, either normal or cancerous. We named this subset the "human colon global database," and we annotated all its samples as normal colon (n = 173), benign colonic adenoma (n = 68) or colorectal cancer (n = 1443). To avoid redundancies (that is, identical samples deposited two or more times in independent GEO data sets) we cross-checked all samples and removed duplicates. When available, we collected all available clinical, pathological and molecular information related to the corresponding patients. As not all arrays were annotated for all variables, individual hypotheses were tested on specific subsets of the "human colon global database." A list of all GEO data sets used in this study, and of their contribution to different experiments, is provided in **Supplementary Table 1**.

Mining of gene-expression arrays using Boolean implications. Gene-expression thresholds between positive and negative samples were defined using the StepMiner algorithm[40], and Boolean implication relationships between pairs of genes using the BooleanNet software[20]. Briefly, for each gene, individual samples were ordered from low-to-high based on their gene-expression values, and a rising step function was fit to the data, trying to minimize differences between fitted and measured values. This method identifies a "step" at the point of largest jump from low to high values (but only if a sufficient number of gene-expression values is present on each side of the jump to exclude a random oscillation due to noise) and sets the gene-expression threshold at the value corresponding to the step[40]. An intermediate region is defined around the threshold, with a width of 1 (threshold +/−0.5), corresponding to a twofold change in expression levels, which represents the minimum noise in these data sets[20,40]. All samples below the intermediate region (< 1st StepMiner threshold − 0.5) are considered negative, and all samples above the intermediate region (> 1st StepMiner threshold + 0.5) are considered positive. When gene-expression levels display a large dynamic range, the StepMiner algorithm can be used to calculate two distinct thresholds: a first threshold to discriminate between "negative" and "positive" samples (1st StepMiner threshold) and a second threshold to split "positive" samples into two subgroups with "low" and "high" gene-expression (2nd StepMiner threshold; **Supplementary Fig. 20**).

We started our search for developmentally regulated genes on the "human colon global database" (**Supplementary Table 1**). To minimize the risk of results being affected by samples containing substantial contaminations from tissues other than colorectal epithelium (e.g., normal liver tissue in hepatic metastases), we restricted our investigation to the subset of arrays with an *EpCAM*$^{+}$/*albumin*$^{-}$ gene-expression profile (**Supplementary Fig. 6**). Threshold gene-expression levels were calculated using the StepMiner algorithm, based on the 1,684 arrays of the "human colon global database" (*EpCAM*$^{+}$: Affymetrix probe 201839_s_at >10.05; *albumin*$^{-}$: Affymetrix probe 211298_s_at <7.97). This operation removed 116 arrays (6.9%) and left 1,568 arrays (93.1%) for analysis (normal colon: n = 170; colorectal adenoma: n = 68; colorectal carcinoma: n = 1,330).

Boolean implication relationships between pairs of genes were systematically computed using the BooleanNet software[20]. Mature enterocyte genes were predicted as genes highly expressed in *KRT20*$^{+}$ arrays and filtered based on the fulfillment of the "*X*$^{+}$ implies *KRT20*$^{+}$" Boolean implication (**Supplementary Fig. 7**). Goblet genes were predicted as genes highly expressed in *MUC2*$^{+}$ arrays and filtered based on the fulfillment of at least one of three independent Boolean implications: "*MUC2* is equivalent to *X*", "*X*$^{+}$ implies *MUC2*$^{+}$", "*MUC2*$^{+}$ implies *X*$^{+}$" (**Supplementary Fig. 8**). Immature genes were predicted as genes highly expressed in *KRT20*$^{-}$ arrays, and filtered based on the fulfillment of the "*KRT20*$^{-}$ implies *X*$^{+}$" Boolean implication (**Supplementary Fig. 9**). Threshold gene-expression levels were calculated using the StepMiner algorithm, based on the global collection of 46,047 human arrays. Gene-expression patterns were considered to fulfill a Boolean implication when the false-discovery rate (FDR) of a sparsity test in the relevant quadrant was <0.05 (ref. 20).

Differences in gene-expression levels among different sample groups (e.g., normal versus adenoma) were evaluated using box plots and tested for statistical significance using a 2-sample t-test (2-tailed). Correlations between two genes' expression levels were measured using Pearson correlation coefficients.

Stratification of human colon cancer patients in distinct gene-expression groups. Associations between gene-expression profiles and patient survival were investigated using a new bioinformatics tool, named Hegemon. Hegemon is an upgrade of the BooleanNet software, where individual gene-expression arrays, after being plotted on a two-axis chart based on the expression of two given genes[20], can be grouped and compared for survival outcomes, using both Kaplan-Meier curves and multivariate analysis based on the Cox proportional hazards method.

Survival analysis was done on a gene-expression database annotated with disease-free survival (DFS) information on 299 patients from three institutions: H. Lee Moffit Cancer Center (n = 164), Vanderbilt Medical Center (n = 55) and Royal Melbourne Hospital (n = 80). This database was created by pooling information from two publicly available and partially redundant GEO data sets (GSE14333, GSE17538; **Supplementary Table 1**)[41,42], both collected on Affymetrix U133 Plus 2.0. To avoid bias due to redundancies (that is, identical samples deposited in both GEO data sets), we cross-checked all samples and removed duplicates.

Guided by single-cell PCR results, we chose to stratify patients using four genes characteristic of top-of-the-crypt *CA1*$^{+}$/*SLC26A3*$^{+}$ cells (*CA1*, *MS4A12*, *CD177*, *SLC26A3*) as markers of terminal differentiation, and using *KRT20*, whose expression is observed in both top-of-the-crypt *CA1*$^{+}$/*SLC26A3*$^{+}$ cells and a subset of *MUC2*$^{+}$/*TFF3*^{high} goblet-type cells, as a more promiscuous marker of both intermediate and terminal differentiation. The hypothesis behind this approach was that, on average, a tumor's overall gene-expression profile would most closely resemble that of the most abundant cell population. Thus, tumors highly enriched in mature, terminally differentiated cell types would be characterized by a lower proliferation rate and/or a lower content of long-term self-renewing cells, and be associated with a better prognosis as compared to tumors predominantly composed by immature, progenitor-like cells.

Threshold gene-expression levels were calculated using the StepMiner algorithm, based on the 25,576 arrays on Affymetrix U133 Plus 2.0. *KRT20* expression (Affymetrix probe 213953_at) was tested as a marker to separate poorly differentiated tumors (*KRT20*$^{-}$) from differentiated ones (*KRT20*$^{+}$). Based on our previous experience[40], we defined as *KRT20*$^{-}$ all tumors whose *KRT20* expression values were < 1st StepMiner threshold − 0.5 (Affymetrix probe 213953_at < 7.00). Genes expressed in top-of-the-crypt *CA1*$^{+}$/*SLC26A3*$^{+}$ cells (*CA1*, *MS4A12*, *CD177*, *SLC26A3*) were tested as markers to separate terminally differentiated tumors (*top-crypt*^{high}) from moderately differentiated ones (*top-crypt*$^{-/low}$). In the case of *CD177* (Affymetrix probe 219669_at) and *SLC26A3* (Affymetrix probes 215657_at), the sensitivity of the probe appeared lower, and its dynamic range narrower, as compared to *CA1* (Affymetrix probe

205950_s_at) or *MS4A12* (Affymetrix probe 220834_at) (**Supplementary Fig. 7**). To maintain consistency in grouping samples with the highest expression levels, we adopted a scaled approach designed to match the different sensitivity of individual gene-expression probes (**Supplementary Fig. 20**). In the case of *CD177* and *SLC26A3*, we chose to separate negative samples from positive ones (*CD177*$^{-}$ versus *CD177*$^{+}$, *SLC26A3*$^{-}$ versus *SLC26A3*$^{+}$), whereas in the case of *CA1* and *MS4A12* we chose to separate low-to-negative expression samples from high expression ones (*CA1*$^{-/low}$ versus *CA1*^{high}, *MS4A12*$^{-/low}$ versus *MS4A12*^{high}). As a result, when we tested *CD177* or *SLC26A3* we defined as *top-crypt*^{high} all tumors that scored as *CD177*$^{+}$ or SLC26A3^{+}, defined as expression values > 1st StepMiner threshold + 0.5 (*CD177*: Affymetrix probe 219669_at > 8.14; *SLC26A3*: Affymetrix probe 215657_at > 5.43), and when we tested *CA1* or *MS4A12* we defined as *top-crypt*^{high} all tumors that scored as *CA1*^{high} or *MS4A12*^{high}, defined as expression values > 2nd StepMiner threshold (*CA1*: Affymetrix probe 205950_s_at > 11.14; *MS4A12*: Affymetrix probe 220834_at > 9.27).

Based on these definitions, we stratified colon tumors into three "gene-expression groups": Group 1 (*KRT20*$^{+}$/*top-crypt*^{high}), Group 2 (*KRT20*$^{+}$/*top-crypt*$^{-/low}$), Group 3 (*KRT20*$^{-}$/*top-crypt*$^{-/low}$). As predicted by the strong Boolean relationship linking *KRT20* to all mature enterocyte genes (**Supplementary Fig. 7**), no tumors were observed that corresponded to the theoretical fourth group (*KRT20*$^{-}$/*top-crypt*^{high}), with the only exception of one isolated sample in the *KRT20/SLC26A3* experiment. In experiments involving comparisons to the EphB2^{+} "intestinal stem cell" (EphB2-ISC) signature (**Supplementary Fig. 23**), tumors were grouped in three categories (EphB2-ISC^{low}, EphB2-ISC^{medium}, EphB2-ISC^{high}), as described in Merlos-Suarez *et al.*[19].

Survival analysis and other statistical tests. Once grouped based on gene-expression profiles, patient subsets were compared for survival outcomes using Kaplan-Meier curves and multivariate analysis based on the Cox proportional hazards method. Differences in Kaplan-Meier curves were tested for statistical significance using the log-rank test. Enrichment of selected pathological or molecular features, such as high pathological grade (G3-G4) or microsatellite instability (MSI), in groups characterized by immature gene-expression patterns (Group-3, *KRT20*$^{-}$/*top-crypt*$^{-/low}$) was measured using odds-ratios and tested for significance using Pearson's χ^2 test.

45. Dalerba, P. *et al.* Reconstitution of human telomerase reverse transcriptase expression rescues colorectal carcinoma cells from in vitro senescence: evidence against immortality as a constitutive trait of tumor cells. *Cancer Res.* **65**, 2321–2329 (2005).
46. Ringner, M. What is principal component analysis? *Nat. Biotechnol.* **26**, 303–304 (2008).
47. O'Doherty, U., Swiggard, W.J. & Malim, M.H. Human immunodeficiency virus type 1 spinoculation enhances infection through virus binding. *J. Virol.* **74**, 10074–10080 (2000).
48. Wang, G.P. *et al.* DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res.* **36**, e49 (2008).
49. Ishizawa, K. *et al.* Tumor-initiating cells are rare in many human tumors. *Cell Stem Cell* **7**, 279–282 (2010).
50. Quintana, E. *et al.* Efficient tumour formation by single human melanoma cells. *Nature* **456**, 593–598 (2008).

doi:10.1038/nbt.2038

Appendix D

Debashis Sahoo. The power of Boolean implication networks. Front. Physio. 23 July 2012, 3:276. doi:10.3389/fphys.2012.00276 (mini review)

frontiers in
PHYSIOLOGY

MINI REVIEW ARTICLE
published: 23 July 2012
doi: 10.3389/fphys.2012.00276

The power of Boolean implication networks

Debashis Sahoo*

Institute of Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA

Edited by:
Hans Westerhoff, University of Manchester, UK

Reviewed by:
Andrzej Michal Kierzek, University of Surrey, UK
Noriko Hiroi, Keio University, Japan
Kristina Gruden, National Institute of Biology, Slovenia

***Correspondence:**
Debashis Sahoo, Institute of Stem Cell Biology and Regenerative Medicine, Stanford University, 265 Campus Drive, Rm G3101B, Stanford, CA, USA.
e-mail: sahoo@stanford.edu

Human diseases have been investigated in the context of single genes as well as complex networks of genes. Though single gene approaches have been extremely successful in the past, most human diseases are complex and better characterized by multiple interacting genes commonly known as networks or pathways. With the advent of high-throughput technologies, a recent trend has been to apply network-based analysis to the huge amount of biological data. Analysis on Boolean implication network is one such technique that distinguishes itself based on its simplicity and robustness. Unlike traditional analyses, Boolean implication networks have the power to break into the mechanistic insights of human diseases. A Boolean implication network is a collection of simple Boolean relationships such as "if A is high then B is low." So far, Boolean implication networks have been employed not only to discover novel markers of differentiation in both normal and cancer tissues, but also to develop robust treatment decisions for cancer patients. Therefore, analyses based on Boolean implication networks have potential to accelerate discoveries in human diseases, suggest therapeutics, and provide robust risk-adapted clinical strategies.

Keywords: bioinformatics, cancer, computational biology, differentiation, microarray analysis, prognostic biomarkers, stem cell, systems biology

INTRODUCTION

In the past detailed single gene investigations in the context of human diseases was extremely successful and produced many useful drugs (Miller et al., 1982; Slamon et al., 2001; Cunningham et al., 2004; Scott et al., 2012). However, the progress was extremely slow and the success was achieved at the cost of a huge number of failed investigations with multiple billions of dollars in investments (Arrowsmith, 2011; Allison, 2012). Unlike in the past years, it is now easy to gather information from tens of thousands of genes simultaneously. Modern approaches can leverage these huge amounts of biological data to understand human diseases. Therefore, a recent trend in analysis has been shifted to multiple genes that are part of a single functional unit commonly known as networks or pathways. The new approaches have been termed network analysis or systems biology. Clearly, these new approaches have the potential to tackle the complexity of human diseases (Mootha et al., 2003; Segal et al., 2003; Basso et al., 2005; Subramanian et al., 2005; Margolin et al., 2006; Bonneau et al., 2007; Lee et al., 2009; Schadt et al., 2010; Bousquet et al., 2011; Gupta et al., 2011; Jornsten et al., 2011). However, the systematic noise in the system has always challenged these approaches. The noise in the system is due to experimental or biological noise and also noise in measuring gene expression values in a microarray hybridization experiment. In addition to noise, other challenge to the network-based approaches is to translate the discoveries to the clinic.

In this mini review, we discuss a systems biology or network-based analysis using Boolean implication network (Sahoo et al., 2008). A Boolean implication network is simply a collection of Boolean implication relationships as described by Sahoo et al. (2008). Boolean typically means a logic calculus of two values, which are high and low gene expression values in this context. A Boolean implication relationship is a simple "if-then" relationship between the high and low gene expression values between a pair of genes. For example, "if A is high, then B is high" is a Boolean implication relationship between a pair of genes A and B, where A high and B low is ruled out as a possible scenario as shown in **Figure 1**. Therefore, whenever gene expression of A is high, we observe gene expression of B is also high. In other words, A high is a subset of B high. In a two dimensional scatter plot between two genes and their thresholds for high and low values, there are four possible quadrants: "A low B low," "A low B high," "A high B low," and "A high B high." One or more sparse quadrants in this plot is mathematically represented as a Boolean implication. For example, the Boolean implication "if A high, then B high" represent a sparse "A high B low" quadrant. There are six possible Boolean implication relationships, two of them are symmetric, and other four are asymmetric. The symmetric Boolean implication relationship has two diagonally opposite sparse quadrant and the asymmetric Boolean implication relationship has only one sparse quadrant. As shown in **Figure 1**, the threshold to define "high" and "low" gene expression levels are determined using StepMiner (Sahoo et al., 2007). The expression levels of each probeset are sorted and a step function fitted (using StepMiner) to the sorted expression level that minimizes the square error between the original and the fitted values. We determined the noise margin by using very tightly correlated genes and found that there is still a difference of twofold change (in log scale a value of Miller et al., 1982) among the values that are linearly related. Therefore, we used a noise margin of 1 (threshold −0.5 to threshold +0.5) and discarded all the microarrays that fall within these region for Boolean implication analysis. The noise margin was an important consideration

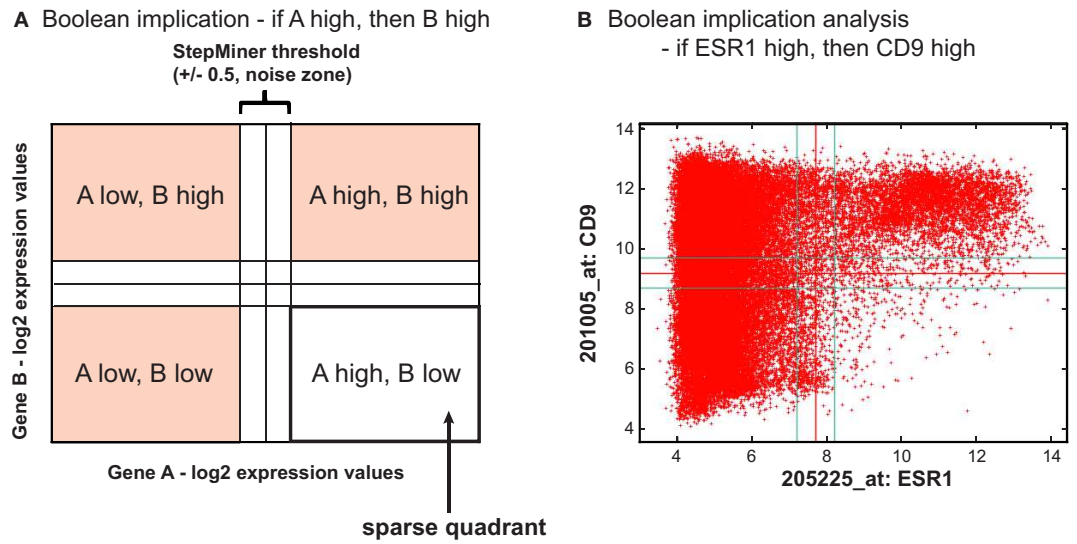


FIGURE 1 | Boolean implication in gene expression database. Boolean implication is a pair-wise gene expression relationship between two genes with respect to their gene expression values. **(A)** Schematic example of a Boolean implication between two genes A and B. Threshold to separate high and low gene expression values are computed using StepMiner. A noise margin of 0.5 is used for statistical calculations. Each of the four quadrant is tested for their sparsity. In this case, A high and B low quadrant is sparse representing the Boolean implication "if A high, then B high." **(B)** An example of a significant Boolean implication between ESR1 and CD9: if ESR1 high, then CD9 high. Every point is a microarray experiment performed on human samples on Affymetrix platform. There are 46,045 microarrays in this scatter plot all of which were downloaded from NCBI's Gene Expression Omnibus (GEO) website.

that allowed us to identify many significant Boolean implication relationships.

SYSTEMS BIOLOGY USING BOOLEAN IMPLICATION

It is possible to discover Boolean implication relationships in the largest possible dataset that include all publicly available microarrays from Gene Expression Omnibus (GEO) or ArrayExpress. These relationships represent natural invariants in a particular species. For example, a Boolean implication relationship in a particular dataset that contains all human samples on Affymetrix platform represents a natural invariant gene expression relationship in human. Many of these invariants are due to tissue specific gene expression. For example, a brain specific gene and a prostate specific gene can never be expressed together. Therefore, they will have a Boolean relationship of the form "if A high, then B low." Similarly, many of these relationships can be due to developmental gene expression pattern or related to the biological process of differentiation. Mining developmentally regulated genes (MiDReG) is a simple algorithm that uses Boolean implication to identify genes expressed at different stages of differentiation (Sahoo et al., 2010). The key concept behind this algorithm is to use invariants to predict state of the gene expression pattern. We describe here how MiDReG and Boolean implication are used in B cell, bladder cancer, and colon cancer differentiation.

B-CELL DIFFERENTIATION

B cells are special types of blood cell that are created from a blood stem cell by the process of differentiation. As the stem cell undergoes the process of differentiation, many genes changes their expression pattern. There are genes that are specific to the stem cell only and also there are genes that are specific to the differentiated B cell. MiDReG algorithm takes advantage of these gene pairs that have a significant Boolean implication "if A high, then B low," and predict other genes that are expressed in the progenitors or precursors of B cells (Inlay et al., 2009; Sahoo et al., 2010). Let's assume that gene A is expressed at the blood stem cells and it turns off as the stem cells differentiate to B cell. Similarly, let's assume that gene B is off at the stem cell and it turns on as the stem cell differentiates to B cells (**Figure 2A**). Therefore, in this narrow view of differentiation gene A and gene B are mutually exclusively expressed. Let's assume that there is a significant Boolean implication "if A high, then B low." The significant Boolean implication represents a global invariant in all microarray datasets. In this case, if we want to identify a gene X that turns on after gene A turns off and before gene B turns on, we could simply use Boolean implication "if A high, X low," and "if B high, X high" (**Figure 2A**). Since the Boolean implication is an invariant, we could hypothesize a state of differentiation where gene A is off, gene X is on, and gene B is off. In addition, this state of differentiation is between stem cell and the mature B cell. Therefore, gene X could potentially mark precursors of the mature B cell. We validated the gene expression patterns of the newly discovered genes using this approach by qPCR on the sorted B-cell progenitors from mouse blood and bone marrow. Review of the published literature of knockout mice revealed that many of our discovered genes were directly involved in B-cell differentiation. Out of 62 MiDReG genes, 41 genes were found to be knocked out in mice. Out of these 41 mice knockouts, 26 (63.4%) genes show defects in B-cell function and differentiation, 9 (22.0%) genes are associated with known B-cell function according to other experiments, and 6

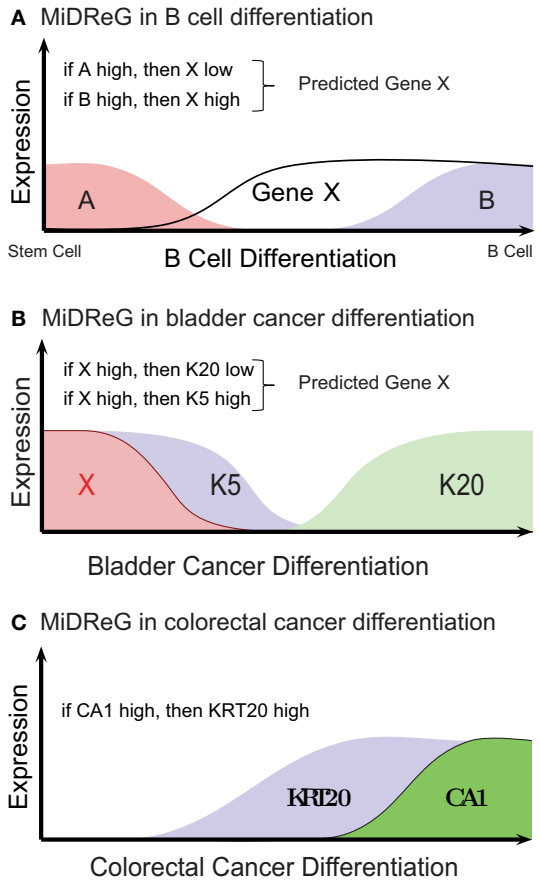


FIGURE 2 | Discovery of markers of differentiation using MiDReG algorithm. Mining developmentally regulated genes (MiDReG) is an algorithm that uses Boolean implication to predict specific markers of differentiation in normal and cancer tissues. **(A)** MiDReG algorithm is used to predict markers of B-cell differentiation. **(B)** MiDReG algorithm is used to predict markers of bladder cancer differentiation. **(C)** MiDReG algorithm is used to predict markers of colorectal cancer differentiation.

(14.6%) genes could have a B-cell function based on their expression in the B cell and reported other hematopoietic functions. A detailed analysis on mouse lineages using MiDReG revealed a new earliest marker of B-cell differentiation Ly6D. This gene was investigated in detail by Inlay et al. (2009). Overall, our results on the B-cell differentiation suggested that MiDReG is a simple but extremely powerful approach to discover novel markers of progenitor cells.

BLADDER CANCER DIFFERENTIATION

Differentiation within cancer is a very controversial topic (Reya et al., 2001). However, in bladder cancer it is established that there are two different cell types identified by Keratin 5 and Keratin 20 (Chan et al., 2009). Keratin 5 marks immature cell types that can differentiate to Keratin 20 positive cells (Chan et al., 2009). MiDReG algorithm was used to identify an upstream marker Keratin 14 (Volkmer et al., 2012). There is a significant Boolean implication relationship between Keratin 5 and Keratin 20 "if Keratin 5 high, then Keratin 20 low" that enabled the MiDReG algorithm to predict upstream markers. In this case, we are interested in a marker X that goes down early compared to Keratin 5. Thus, it is expressed at the most immature state of the cancer cell. The candidate markers were chosen based on Boolean implication "if X high, then Keratin 5 high" and "if X high, then Keratin 20 low" (**Figure 2B**). Keratin 14 was one of the markers that satisfied these two Boolean implication strongly. In addition, Keratin 14 was a single prognostic marker in both gene and protein expression datasets. The prognostic power of Keratin 14 was independent of currently established stage and grade. Therefore, a simple immunohistochemical analysis can identify high risk bladder cancer patients. Since, clinicians decide whether to perform cystectomy which is complete bladder removal based on stage and grade, it is possible to incorporate Keratin 14 based risk stratification into this important clinical decision endpoint. Clinicians are currently developing risk-adapted clinical strategies based on Keratin 14 for bladder cancer patients.

COLON CANCER DIFFERENTIATION

Many important markers in the differentiation of colon cancer cells follow Boolean implication (Dalerba et al., 2011). For example, there is a significant Boolean implication between Keratin 20 and CA1 "if CA1 high, then Keratin 20 high" (**Figure 2C**). This relationship is particularly strong with no exception. There are no tumors with CA1 high and KRT20 low. Even in a tumor when CA1 positive cells are present they have to go through a KRT20 positive precursor cell during differentiation. Accordingly, CA1 positive cells are a subset of Keratin 20 positive cells in both normal colon and colorectal cancer tissues. In addition, Keratin 20 negative patients have worse outcome compared to CA1 positive and Keratin 20 positive cancer patients. Other markers such as MS4A12, CD177, and SLC26A3 follow similar Boolean implication relationships.

STRENGTHS AND LIMITATIONS

In this review we show that Boolean implication can be used to identify markers of differentiation in both normal and cancer tissues. The strength of Boolean implication is its ability to identify asymmetric gene expression relationships. In contrast, most other approaches focus on using symmetric gene expression relationship to build gene expression network. We have shown that some of the gene expression patterns in differentiation can be modeled using asymmetric Boolean implication. Therefore, it would be useful for predicting important genes involved in the process of differentiation. In addition, markers of differentiation are most likely robust prognostic biomarkers in cancer patients. Using these markers, clinicians may be able to develop better risk-adapted treatment decisions for cancer patients. The limitation of Boolean implication is that it requires large number of samples. Also, it might miss many other important genes that are involved in differentiation but do not have significant Boolean implication. Accordingly, Boolean implication is a very stringent criterion. Therefore, it pulls out many important genes and appears to be less noisy compared to traditional approaches.

An important distinction between Boolean implication analyses compared to other traditional network-based analyses is that most of these other analyses are focused on identifying gene regulatory networks or signal transduction pathways. Boolean

implication has not been utilized to identify gene regulatory networks or signaling networks which contains simple feed-back and feed-forward structure. Instead, it was used to identify cell type or tissue specific gene expression patterns and they are interpreted in terms of development and differentiation. This is very different from Bayesian or mutual information based networks that primarily identify transcription factors and their targets (Segal et al., 2003; Basso et al., 2005; Margolin et al., 2006; Lee et al., 2009). Similarly, Boolean implication analyses are also different from traditional Boolean networks that are used to build a functional executable model or a circuit model (Glass and Kauffman, 1973; Shmulevich and Kauffman, 2004). There are also networks based on ODE models which describes mechanistic biochemical interactions (Ferrell et al., 2011). Both the Boolean and ODE based approaches described above models non-linear dynamical systems (Glass and Kauffman, 1973; Shmulevich and Kauffman, 2004; Ferrell et al., 2011). In contrast, Boolean implication analyses models static invariant relationships in a large biological dataset.

In summary, Boolean implication is an empirically observed relationship in the data, which may not hold for data gathered for different tissue types or under different conditions. Like correlation networks, Boolean implication networks do not capture causality. Boolean implication captures both symmetric as well as asymmetric relationships. It provides a powerful platform for discovery of novel markers of differentiation in both normal and cancer tissues.

ACKNOWLEDGMENTS

Boolean implication and MiDReG tools were developed as part of Dr. Sahoo's Ph.D. at Stanford University with significant contribution from Prof. David Dill as Ph.D. advisor, Prof. Sylvia Plevritis as co-advisor, Prof. Rob Tibshirani and Andrew Gentles. The application of these tools were developed in collaboration with the Weissman lab and the Clarke lab at Stanford University. The author thank I. L. Weissman, M. F. Clarke, J. Lipsick, M. van de Rijn, L. D. Shortliffe, J. D. Brooks, J. Pollack, R. Levy, J. Seita, M. Inlay, D. Bhattacharya, R. K. Chin, J. Volkmer, P. Dalerba, K. S. Chan for critical discussions, helpful suggestions, and technical advice. Dr. Sahoo is supported by National Institutes of Health (NIH) Grant K99CA151673-01A1, Department of Defense Grant W81XWH-10-1-0500, Ludwig Institute Grant (PI: Irv Weissman), and a grant from the Siebel Stem Cell Institute and the Thomas and Stacey Siebel Foundation. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH and other grant agencies.

REFERENCES

Allison, M. (2012). Reinventing clinical trials. *Nat. Biotechnol.* 30, 41–49.

Arrowsmith, J. (2011). Trial watch: phase III and submission failures: 2007–2010. *Nat. Rev. Drug Discov.* 10, 87.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat. Genet.* 37, 382–390.

Bonneau, R., Facciotti, M. T., Reiss, D. J., Schmid, A. K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M. H., Bare, J. C., Longabaugh, W., Vuthoori, M., Whitehead, K., Madar, A., Suzuki, L., Mori, T., Chang, D. E., Diruggiero, J., Johnson, C. H., Hood, L., and Baliga, N. S. (2007). A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131, 1354–1365.

Bousquet, J., Anto, J. M., Sterk, P. J., Adcock, I. M., Chung, K. F., Roca, J., Agusti, A., Brightling, C., Cambon-Thomsen, A., Cesario, A., Abdelhak, S., Antonarakis, S. E., Avignon, A., Ballabio, A., Baraldi, E., Baranov, A., Bieber, T., Bockaert, J., Brahmachari, S., Brambilla, C., Bringer, J., Dauzat, M., Ernberg, I., Fabbri, L., Froguel, P., Galas, D., Gojobori, T., Hunter, P., Jorgensen, C., Kauffmann, F., Kourilsky, P., Kowalski, M. L., Lancet, D., Pen, C. L., Mallet, J., Mayosi, B., Mercier, J., Metspalu, A., Nadeau, J. H., Ninot, G., Noble, D., Oztürk, M., Palkonen, S., Préfaut, C., Rabe, K., Renard, E., Roberts, R. G., Samolinski, B., Schünemann, H. J., Simon, H. U., Soares, M. B., Superti-Furga, G., Tegner, J., Verjovski-Almeida, S., Wellstead, P., Wolkenhauer, O., Wouters, E., Balling, R., Brookes, A. J., Charron, D., Pison, C., Chen, Z., Hood, L., and Auffray, C. (2011). Systems medicine and integrated care to combat chronic noncommunicable diseases. *Genome Med.* 3, 43.

Chan, K. S., Espinosa, I., Chao, M., Wong, D., Ailles, L., Diehn, M., Gill, H., Presti, J. Jr., Chang, H. Y., van de Rijn, M., Shortliffe, L., and Weissman, I. L. (2009). Identification, molecular characterization, clinical prognosis, and therapeutic targeting of human bladder tumor-initiating cells. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14016–14021.

Cunningham, D., Humblet, Y., Siena, S., Khayat, D., Bleiberg, H., Santoro, A., Bets, D., Mueser, M., Harstrick, A., Verslype, C., Chau, I., and Van Cutsem, E. (2004). Cetuximab monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *N. Engl. J. Med.* 35, 337–345.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. F., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. F., and Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* 29, 1120–1127.

Ferrell, J. E., Tsai, T. Y., and Yang, Q. (2011). Modeling the cell cycle: why do certain circuits oscillate? *Cell* 144, 874–885.

Glass, L., and Kauffman, S. A. (1973). The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* 39, 103–129.

Gupta, P. B., Fillmore, C. M., Jiang, G., Shapira, S. D., Tao, K., Kuperwasser, C., and Lander, E. S. (2011). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell* 146, 633–644.

Inlay, M. A., Bhattacharya, D., Sahoo, D., Serwold, T., Seita, J., Karsunky, H., Plevritis, S. K., Dill, D. L., and Weissman, I. L. (2009). Ly6d marks the earliest stage of B-cell specification and identifies the branchpoint between B-cell and T-cell development. *Genes Dev.* 23, 2376–2381.

Jornsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., and Nelander, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.* 7, 486. doi:10.1038/msb.2011.17

Lee, S. I., Dudley, A. M., Drubin, D., Silver, P. A., Krogan, N. J., Pe'er, D., and Koller, D. (2009). Learning a prior on regulatory potential from eQTL data. *PLoS Genet.* 5, e1000358. doi:10.1371/journal.pgen.1000358

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl. 1), S7.

Miller, R. A., Maloney, D. G., Warnke, R., and Levy, R. (1982). Treatment of B-cell lymphoma with monoclonal anti-idiotype antibody. *N. Engl. J. Med.* 306, 517–522.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273.

Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature* 414, 105–111.

Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani, R., and Plevritis, S. K. (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biol.* 9, R157.

Sahoo, D., Dill, D. L., Tibshirani, R., and Plevritis, S. K. (2007). Extracting binary signals from microarray time-course data. *Nucleic Acids Res.* 35, 3705–3712.

Sahoo, D., Seita, J., Bhattacharya, D., Inlay, M. A., Weissman, I. L., Plevritis, S. K., and Dill, D. L. (2010). MiDReG: a method of mining developmentally regulated genes using Boolean implications. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5732–5737.

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., and Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* 11, 647–657.

Scott, A. M., Wolchok, J. D., and Old, L. J. (2012). Antibody therapy of cancer. *Nat. Rev. Cancer* 12, 278–287.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176.

Shmulevich, I., and Kauffman, S. A. (2004). Activities and sensitivities in Boolean network models. *Phys. Rev. Lett.* 93, 048701.

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., and Norton, L. (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* 344, 783–792.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

Volkmer, J. P., Sahoo, D., Chin, R. K., Ho, P. L., Tang, C., Kurtova, A. V., Willingham, S. B., Pazhanisamy, S. K., Contreras-Trujillo, H., Storm, T. A., Lotan, Y., Beck, A. H., Chung, B. I., Alizadeh, A. A., Godoy, G., Lerner, S. P., van de Rijn, M., Shortliffe, L. D., Weissman, I. L., and Chan, K. S. (2012). Three differentiation states risk-stratify bladder cancer into distinct subtypes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2078–2083.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 March 2012; paper pending published: 04 April 2012; accepted: 27 June 2012; published online: 23 July 2012.
Citation: Sahoo D (2012) The power of Boolean implication networks. Front. Physio. ***3****:276. doi: 10.3389/fphys.2012.00276*
This article was submitted to Frontiers in Systems Physiology, a specialty of Frontiers in Physiology.

Appendix E

NIH pathway to independence award (K99/R00) award

CAREER TRANSITION AWARD
Department of Health and Human Services
National Institutes of Health
NATIONAL CANCER INSTITUTE

Issue Date: 07/01/2011

Grant Number: 1K99CA151673-01A1

Principal Investigator(s):
Debashis Sahoo, PHD

Project Title: Application of Boolean Networks to discover stem and progenitor cells

Karen Wong
Research Process Manager
301 Ravenswood Avenue
Menlo Park, CA 940253434

Award e-mailed to: NIHAWARDS@lists.stanford.edu

Budget Period: 07/01/2011 – 06/30/2012
Project Period: 07/01/2011 – 06/30/2013

Dear Business Official:

The National Institutes of Health hereby awards a grant in the amount of $152,135 (see “Award Calculation” in Section I and “Terms and Conditions” in Section III) to STANFORD UNIVERSITY in support of the above referenced project. This award is pursuant to the authority of 42 USC 241, 42 CFR 52, 42 CFR 67 and is subject to the requirements of this statute and regulation and of other referenced, incorporated or attached terms and conditions.

Acceptance of this award including the “Terms and Conditions” is acknowledged by the grantee when funds are drawn down or otherwise obtained from the grant payment system.

Each publication, press release or other document that cites results from NIH grant-supported research must include an acknowledgment of NIH grant support and disclaimer such as “The project described was supported by Award Number K99CA151673 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.”

Award recipients are required to comply with the NIH Public Access Policy. This includes submission to PubMed Central (PMC), upon acceptance for publication, an electronic version of a final peer-reviewed, manuscript resulting from research supported in whole or in part, with direct costs from National Institutes of Health. The author's final peer-reviewed manuscript is defined as the final version accepted for journal publication, and includes all modifications from the publishing peer review process. For additional information, please visit http://publicaccess.nih.gov/.

Award recipients must promote objectivity in research by establishing standards to ensure that the design, conduct and reporting of research funded under NIH-funded awards are not biased by a conflicting financial interest of an Investigator. Investigator is defined as the Principal Investigator and any other person who is responsible for the design, conduct, or reporting of NIH-funded research or proposed research, including the Investigator's spouse and dependent children. Awardees must have a written administrative process to identify and manage financial conflict of interest and must inform Investigators of the conflict of interest policy and of the Investigators' responsibilities. Prior to expenditure of these awarded funds, the Awardee must report to the NIH Awarding Component the existence of a conflicting interest and within 60 days of any new conflicting interests identified after the initial report. Awardees must comply with these and all other aspects of 42 CFR Part 50, Subpart F. These requirements also apply to subgrantees, contractors, or collaborators engaged by the Awardee under this award. The NIH website http://grants.nih.gov/grants/policy/coi/index.htm provides additional information.

If you have any questions about this award, please contact the individual(s) referenced in Section IV.

Sincerely yours,

Amy Connolly
Grants Management Officer
NATIONAL CANCER INSTITUTE

Additional information follows

Supplmentary Figure 1: List of prostate cancer datasets

A. Prostate Cancer datasets

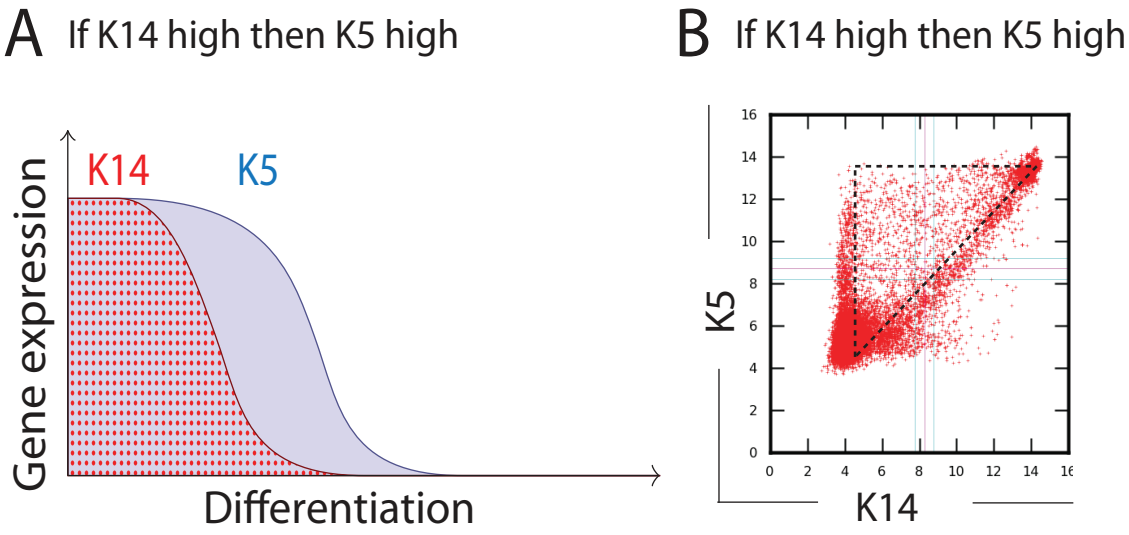
Name	Journal	Year	Pubmed	RAW	GEO/AE	Platform	Survival	# patients
Singh D	Cancer Cell	2002	12086878	yes	NA	U95Av2	no	102
Glinsky GV	J Clin Invest.	2004	15067324	yes	NA	U133A2	yes	79
Lapointe J	PNAS	2004	14711987	yes	GSE3933	cDNA	yes	112
Zuls	Genome Res	2010	21521786	yes	NA	HEEBO	yes	131
Chandran UR	BMC Cancer	2007	17430594		GSE6919	HG_U95Av2	no	171
Pressinotti NC	Mol Cancer	2010	20035634		GSE15484	GPL3050	no	65
Sboner A	BMC Med Genomics	2010	20233430	yes	GSE16560	GPL5474	yes	281
Wang Y	Cancer Res	2009	20663908	yes	GSE17951	U133Plus2	no	154
Taylor BS	Cancer Cell	2010	20579941	yes	GSE21034	HuEx-1_0-st	yes	367
Setlur	J Natl Cancer Inst	2008	18505969	yes	GSE8402	GPL5474	no	472

B. Global Prostate Cancer Database

Name	Journal	Year	Pubmed	RAW	GEO/AE	Platform	Survival	# patients
Bakshi S	Environ Health Perspect	2008	18560533	yes	GSE9951	GPL570		19
Berry PA	Prostate	2011	21432868	yes	E-MTAB-402	GPL570		14
Best CJ	Clin Cancer Res	2005	16203770	yes	GSE2443	GPL96		20
Birnie R	Genome Biol.	2008	18492237	yes	E-MEXP-993	GPL570		36
Chambers KF	J Biomed Sci	2011	21696611	yes	E-MEXP-2034	GPL570		40
Guyon I		2011		yes	E-TABM-456	GPL96		85
Liu P	Cancer Res	2006	16618720	yes	E-TABM-26	GPL96		57
Sun Y	Prostate	2009	19343730	yes	GSE25136	GPL96		79
Traka M	PLoS One.	2008	18596959	yes	E-MEXP-1243	GPL570		81
Tsavachidou D	J Natl Cancer Inst.	2009	19244175	yes	E-MEXP-1327	GPL96		85
Varambally S	Cancer Cell	2005	16286247	yes	GSE3325	GPL570		19
Wallace TA	Cancer Res	2008	18245496	yes	GSE6956	GPL571		72
Wang Y	Cancer Res	2010	20663908	yes	GSE8218	GPL96		130
Wang Y	Cancer Res	2010	20663908	yes	GSE17951	GPL570		154
							Total	891

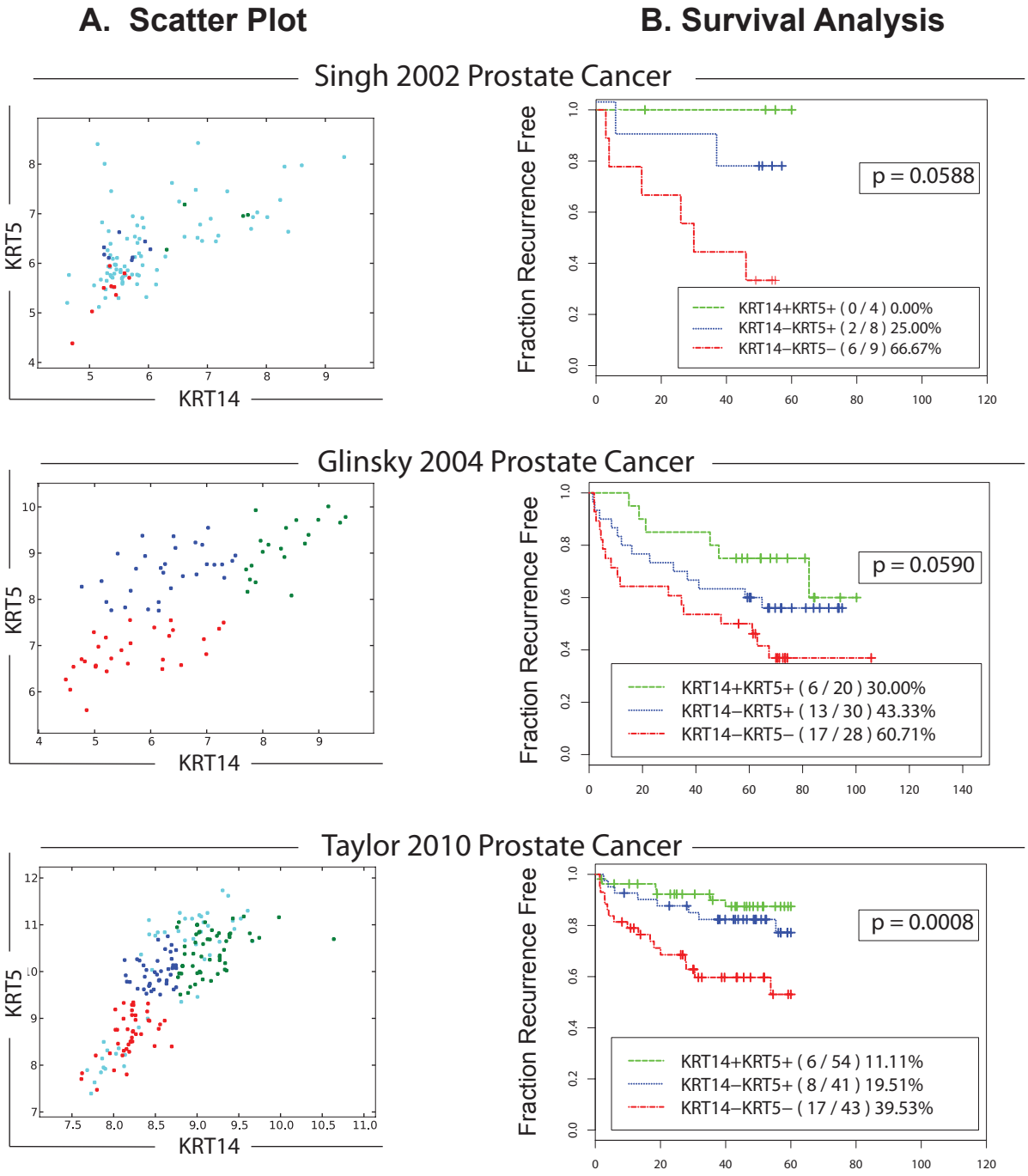
Supplementary Figure 1: List of prostate cancer datasets. Panel A shows a list of publicly available prostate cancer datasets with clinical information (Only five dataset with survival outcome). Panel B shows a list of prostate cancer datasets on Affymetrix U133A (GPL96), U133A 2.0 (GPL571) or U133 Plus 2.0 (GPL570) microarray platforms that are normalized together to build a large global prostate cancer database. The lists include the first author, journal where it was first published, year in which it was published, the PubMed id, GEO/ArrayExpress id, microarray platforms, survival annotation, and number of patients.

Supplementary Figure 2: Infering developmental gene regulation from Boolean implication relationship



Supplementary Figure 2: Infering developmental gene regulation from Boolean implication relationship. To infer developmental gene regulation (A) we use Boolean implication (B). In most human epithelial tissues both Keratin 5 (K5) and Keratin 14 (K14) are expressed in the basal cell compartments. We analyzed gene expression values of K14 and K5, that is presented in the form of a scatterplot with 25,237 points representing diverse microarrays on human samples including different normal and cancer tissues. We summarize the gene expression relationship between K14 and K5 as “if K14 high then K5 high” or alternatively a Boolean implication relationship “K14 high => K5 high”. The relationship clearly suggests that K14+ arrays are a subset of K5+ arrays. Since not all cells within a sample express K14 and K5, we could hypothesize that K14+ cells are a subset of K5+ cells (A) based on the Boolean implication. Panel A shows a likely model of developmental gene regulation between K14 and K5, where K14 is upstream of K5.

Supplementary Figure 3: Relationship between Keratin gene expression and clinical outcome



Supplementary Figure 3: **Relationship between Keratin gene expression and clinical outcome.** To evaluate whether Keratin gene expression is associated with patient outcome, we investigated the status of three Keratin expression groups (KRT14+KRT5+, KRT14-KRT5+, KRT14-KRT5-) on recurrence-free survival (RFS) in three independent prostate cancer cohorts (Singh 2002 dataset, n=102; Glinsky 2004 dataset, n=78; Taylor 2010 dataset, n=185), The results confirmed that KRT14-KRT5- tumors were associated with worse clinical outcomes (B). In addition, KRT14+KRT5+ tumors were associated with best clinical and KRT14-KRT5+ tumors were associated with intermediate clinical outcome.